



UNIVERSIDADE ESTADUAL PAULISTA
“JÚLIO DE MESQUITA FILHO”
Câmpus de Presidente Prudente

JUAN GARCIA RUIZ

**APLICAÇÃO DE MODELOS DE REGRESSÃO LOGÍSTICA PARA OS
DADOS PÚBLICOS DA EMPRESA OLIST**

PRESIDENTE PRUDENTE

2023

JUAN GARCIA RUIZ

**APLICAÇÃO DE MODELOS DE REGRESSÃO LOGÍSTICA PARA OS
DADOS PÚBLICOS DA EMPRESA OLIST**

Relatório Final do Trabalho de Conclusão de
Curso apresentado ao Curso de Graduação
em Estatística da FCT/Unesp para
aproveitamento na disciplina Trabalho de
Conclusão de Curso. Orientador: Prof. Dr.
Mário Hissamitsu Tarumoto.

PRESIDENTE PRUDENTE

2023

R934a	Ruiz, Juan Garcia Aplicação de modelos de Regressão Logística para os dados públicos da empresa Olist / Juan Garcia Ruiz. -- Presidente Prudente, 2023 86 p. : il., tabs. Trabalho de conclusão de curso (Bacharelado - Estatística) - Universidade Estadual Paulista (Unesp), Faculdade de Ciências e Tecnologia, Presidente Prudente Orientador: Mário Hissamitsu Tarumoto 1. Análise de Regressão Logística. 2. previsão de vendas. 3. vendedores. 4. comércio eletrônico. 5. base de dados da web. I. Título.
-------	--

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Ciências e Tecnologia, Presidente Prudente. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

TERMO DE APROVAÇÃO

Juan Garcia Ruiz

APLICAÇÃO DE MODELOS DE REGRESSÃO LOGÍSTICA PARA OS DADOS PÚBLICOS DA EMPRESA OLIST

Relatório de Final de Trabalho de Conclusão de Curso aprovado como requisito para obtenção de créditos na disciplina Trabalho de Conclusão do curso de graduação em Estatística da Faculdade de Ciências e Tecnologia da Unesp, pela seguinte banca examinadora:

Orientador: 

Prof. Dr. Mário Hissamitsu Tarumoto
Departamento de Estatística



Prof. Dr. Manoel Ivanildo Silvestre Bezerra
Departamento de Estatística



Prof. Dr. Fernando Antonio Moala
Departamento de Estatística

Presidente Prudente, 27 de janeiro de 2023.

AGRADECIMENTOS

Gostaria de iniciar esta etapa de agradecimentos agradecendo a minha mãe (Christina Palau) e ao meu pai (Clovis Ruiz) por tudo que fizeram por mim ao longo de todos esses anos. Agradeço também a minha família por todo suporte ao longo desses cinco anos. Um agradecimento especial a Amanda Vasconcelos por ser minha companheira durante praticamente toda a jornada na faculdade. Gostaria de agradecer todos os docentes do curso de estatística pelo aprendizado, especialmente ao Prof. Dr. Mário Tarumoto por todo suporte desde o primeiro ano e pela parceria durante este trabalho, agradeço também a banca composta por Prof. Dr. Fernando Moala e Prof. Dr. Manoel Bezerra. E finalmente gostaria de agradecer a uma grande amiga, grande líder e mentora, a Jedi Adriana Silva, sem ela este trabalho seria apenas mais um trabalho. Dri, agradeço imensamente tudo, desde o acolhimento na faculdade, os ensinamentos na empresa e as broncas durante todo o meu processo de aprendizado.

Também agradeço a Deus por tudo e por me guiar e acompanhar durante essa trajetória.

EPÍGRAFE

*“Lo que brilla con luz propia nadie lo puede apagar
Su brillo puede alcanzar la oscuridad de otras cosas”*

~Pablo Milanés (Canción Por La Unidad Latinoamericana)

RESUMO

Com o crescimento acelerado das empresas de varejo por conta do mercado virtual, empresas como a Olist que focam em aproximar pequenos vendedores dos consumidores finais, a cada dia mais vêm ganhando espaço no mercado. Para a Olist, que tem parte do seu faturamento baseado na quantidade de vendas dos vendedores, é interessante saber quais vendedores não estão vendendo. Portanto o objetivo deste trabalho é identificar as características dos vendedores que não efetuarão venda. Na estatística vemos muitas técnicas aprofundadas e de difícil compreensão para pessoas que não são dessa área, portanto um objetivo secundário deste trabalho é ser o mais claro e simples, para que seja de fácil entendimento até para uma pessoa que não seja da área. Ao desenvolver o trabalho aprenderemos sobre o modelo de Regressão Logística, como interpretá-lo, para que serve um modelo, métricas que avaliam sua qualidade, etc. Este trabalho é um trabalho de ciência de dados, ou seja, não contempla apenas os conteúdos aprendidos na faculdade de estatística, mas sim as técnicas e estratégias aprendidas no mercado de trabalho que potencializam a estatística para resolução de problemas de forma criativa. Como resultado da análise, observou-se que o modelo de regressão logística teve um ótimo desempenho no quesito previsão e poderia ser usado para a empresa.

Palavras-chave: Análise de Regressão Logística, previsão de vendas, vendedores, comércio eletrônico, base de dados da Web.

ABSTRACT

With the accelerated growth of the retail companies because of the virtual market, companies like Olist that are focused on approximate small businesses and the final consumers, are getting more space in the market every day. To the Olist, that has part of its income based on the quantity of the seller's sales, it is interesting to know which sellers are not selling. Therefore the goal of this work is to identify the characteristics of the sellers that will sell. In the statistics we see a lot of techniques that are hard to comprehend for people that are not from this area, thus a second goal of this work is to be the most clear and simple that it possibly can, so it will be easily understood even by a person that is not from the area. Throughout the development of the work we learned about the model of Logistic Regression, how to interpret it, for what the model works, metrics that evaluate its quality, etcetera. This work is a data science work, which means that it doesn't behold only the contents learned in college, but the techniques and strategies learned in the workspace that maximize the statistics to solve problems in a creative way. As a result of the analysis, it was observed that the logistic regression model performed well in terms of forecasting and could be used for the company.

Keywords: Logistic Regression Analysis, sales forecast, sellers, e-commerce, web database.

LISTA DE FIGURAS

Figura 1 – Esquema de relacionamento OLIST	13
Figura 2 - Esquema de organização das bases.....	14
Figura 3 - Matriz de confusão.	24
Figura 4 - Aplicação de safras nos dados.....	27
Figura 5 - Processo de modelagem automatizado.....	32
Figura 6 – Objetivo do uso do gráfico de dispersão.....	36
Gráfico 1 – Gráfico de barras da variável fl_venda.....	37
Gráfico 2 – Gráfico de dispersão das variáveis proporção ativação meses e venda.....	37
Gráfico 3 – Boxplot da proporção ativação meses por venda.....	38
Gráfico 4 – Gráfico de dispersão das variáveis dias desde a última venda e venda.....	39
Gráfico 5 – Boxplot dias desde a última venda por venda.....	39
Gráfico 6 – Gráfico de dispersão das variáveis quantidade de produtos no último mês e venda.....	40
Gráfico 7 – Boxplot quantidade de produtos no último mês por venda.....	41
Gráfico 8 – Gráfico de dispersão das variáveis media dias entre dt compra dt entrega últimos 4 meses e venda.....	41
Gráfico 9 – Boxplot media dias entre dt compra dt entrega últimos 4 meses por venda.....	42
Gráfico 10 – Gráfico de dispersão das variáveis quantidade de pedidos cancelados até hoje e venda.....	43
Gráfico 11 – Boxplot quantidade de pedidos cancelados até hoje por venda.....	43
Gráfico 12 – Gráfico de dispersão das variáveis quantidade de pedidos cancelados último mes e venda.....	44
Gráfico 13 – Boxplot quantidade de pedidos cancelados último mês por venda.....	45
Gráfico 14 –Gráfico de dispersão das variáveis quantidade de pedidos outros status até hoje e venda.....	46
Gráfico 15 – Boxplot quantidade de pedidos outros status até hoje por venda.....	46
Gráfico 16 –Gráfico de dispersão das variáveis quantidade de pedidos outros status último mês e venda.....	47
Gráfico 17 – Boxplot quantidade de pedidos outros status último mês por venda.....	48
Gráfico 18 –Gráfico de dispersão das variáveis media qtd parcelas pedidos pgto cc últimos 6 meses e venda.....	49

Gráfico 19 – Boxplot media qtd parcelas pedidos pgto cc últimos 6 meses por venda	49
Gráfico 20 –Gráfico de dispersão das variáveis média do tamanho da descrição por produto e venda	50
Gráfico 21 – Boxplot média do tamanho da descrição por produto por venda	51
Gráfico 22–Gráfico de dispersão das quantidades de categorias vendendo e venda	51
Gráfico 23 – Boxplot quantidade de categorias vendendo por venda	52
Gráfico 24 –Gráfico de dispersão das variáveis quantidade de review nos últimos 3 meses e venda	53
Gráfico 25 – Boxplot quantidade de review nos últimos três meses por venda	53
Gráfico 26 –Gráfico de dispersão das variáveis média das notas de review nos últimos três meses e venda	54
Gráfico 27 – Boxplot média das notas de review nos últimos três meses por venda	55
Figura 7 - Matriz de confusão treino	57
Gráfico 28 – Curva ROC treino	57
Figura 8 - Matriz de confusão teste	67
Gráfico 29 – Curva ROC teste	67

LISTA DE TABELAS

Tabela 1 - Quebras das variáveis.....	55
Tabela 2 – Modelo completo.....	58
Tabela 3 – Modelo Final.....	60
Tabela 4 – Intervalo de confiança dos parâmetros.....	61
Tabela 5 – Exemplo do desempenho do modelo.....	62
Tabela 6 – Interpretação das odds ratio.....	64
Quadro 1 - Dicionário da base olist_customers_dataset.....	72
Quadro 2 - Dicionário da base olist_geolocation_dataset.....	72
Quadro 3 - Dicionário da base olist_order_items_dataset.....	72
Quadro 4 - Dicionário da base olist_order_payments_dataset.....	73
Quadro 5 - Dicionário da base olist_order_reviews_dataset.....	73
Quadro 6 - Dicionário da base olist_orders_dataset.....	73
Quadro 7 - Dicionário da base olist_products_dataset.....	74
Quadro 8 - Dicionário da base olist_sellers_dataset.....	74
Quadro 9 - Dicionário da ABT.....	75

SUMÁRIO

1 INTRODUÇÃO	12
2 EMPRESA OLIST	16
3 REGRESSÃO LOGÍSTICA	17
3.1 MODELO DE REGRESSÃO LOGÍSTICA	17
3.2 ESTIMAÇÃO DOS PARÂMETROS β_j	18
3.3 TESTE DA RAZÃO DE VEROSSIMILHANÇA E TESTE Z-WALD	19
3.4 INTERVALO DE CONFIANÇA DOS PARÂMETROS	20
3.5 DEVIANCE RESIDUAL	20
3.6 ODDS RATIO	21
4 MÉTRICAS DE QUALIDADE DE AJUSTE	23
4.1 OBJETIVO DO MODELO	23
4.2 MATRIZ DE CONFUSÃO	23
4.3 LIFT	25
4.4 MÉTRICAS ASSOCIADAS AO ERRO	25
5 TRATAMENTO DOS DADOS	27
5.1 SAFRAS	27
5.2 DADOS FALTANTES	28
6 ADAPTAÇÃO DA METODOLOGIA CRISP-DM	30
7 DESENVOLVIMENTO	34
7.1 REUNIÃO COM A ÁREA DE NEGÓCIO	34
7.2 CRIAÇÃO DA ABT	35
7.3 TRATAMENTO DOS DADOS	35
7.4 ANÁLISE DESCRITIVA	36
7.5 VARIÁVEIS DUMMY	55
7.6 QUALIDADE DO AJUSTE	56
7.7 MODELO DE REGRESSÃO LOGÍSTICA	58
7.8 INTERPRETAÇÃO DO MODELO	63
7.9 QUALIDADE DA PREVISÃO	66
8 CONSIDERAÇÕES FINAIS	69

REFERÊNCIAS.....	70
APÊNDICES	72
APÊNDICE 1 –DICIONÁRIOS DOS DADOS OBTIDOS NA PLATAFORMA KEAGGLE	72
APÊNDICE 2 –DICIONÁRIO DOS DADOS DA ABT	75
APÊNDICE 3 – LINK E QR-CODE.....	86

1 INTRODUÇÃO

Algumas empresas do ramo de varejo, conhecidas como marketplaces¹, trabalham em um formato bem interessante e que atualmente está fazendo muito sucesso, no qual a empresa disponibiliza espaço para que vendedores parceiros possam anunciar produtos em seu e-commerce². Este formato possibilita que tanto pessoas físicas, quanto outras empresas, façam o papel de vendedores parceiros, ou seja, pode-se comprar um produto em determinado marketplace que está sendo vendido por outra pessoa ou empresa.

A partir desse modelo de negócio, surgiram empresas que focam em gerar visibilidade para pequenos empreendedores e facilitar a inserção dos mesmos em grandes marketplaces. A grande vantagem desta forma de trabalho, é a possibilidade de anunciar em grandes marketplaces, pois gerenciar vários ambientes ao mesmo tempo não é uma tarefa fácil e, além disso, os pequenos empreendedores normalmente não tem visibilidade em grandes plataformas. O Olist³ é uma empresa que identificou este problema e gerou uma solução acessível e viável. Dalvi (apud Fonseca, (2021)), comenta que o Olist surgiu em 2015, com a ideia de ser a maior loja de departamentos dentro das lojas de departamento.

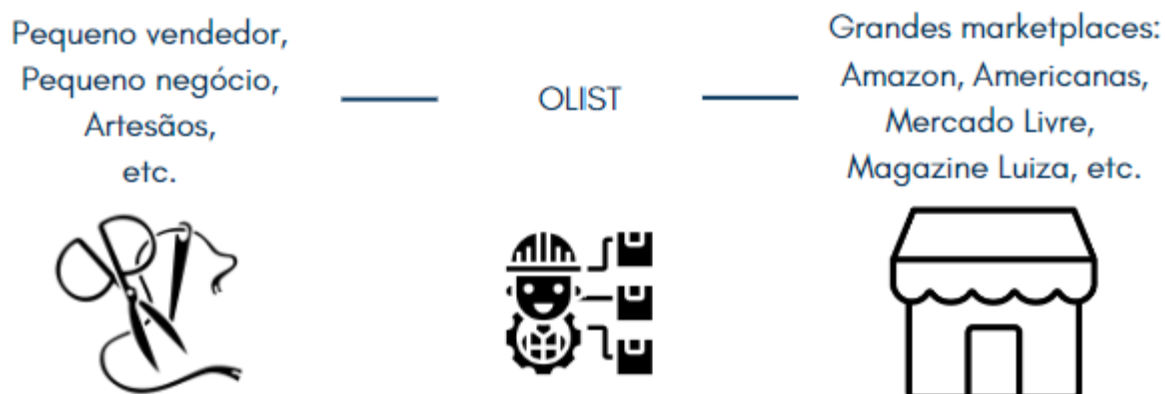
Com este formato de sistema de venda, introduzido anteriormente, os pequenos vendedores têm mais visibilidade, pois podem anunciar com o nome de renomadas empresas (como a Olist), que por sua vez, anunciam seus produtos em diversos marketplaces. Para a empresa que controla a plataforma (marketplaces) esta ação é vantajosa pois pode cobrar taxas sobre o que está sendo vendido, para empresas parceiras, ou seja, a vantagem é cobrar por cada produto vendido e para os pequenos empreendedores as vantagens são: gerenciar tudo da mesma plataforma e estar inserido dentro dos maiores marketplaces com maior visibilidade. O esquema de relacionamento da OLIST com as empresas é ilustrado na Figura 1.

¹ Marketplaces são as plataformas onde os produtos são vendidos. Alguns exemplos de marketplaces são: Amazon, Americanas, Mercado Livre, etc.

² E-commerce é o comércio virtual, ou seja, o e-commerce é composto por todas as empresas que vendem seus produtos virtualmente.

³ Olist é uma empresa de departamentos e tem como principal objetivo facilitar pequenos vendedores a se inserir em grandes lojas virtuais.

Figura 1 – Esquema de relacionamento OLIST

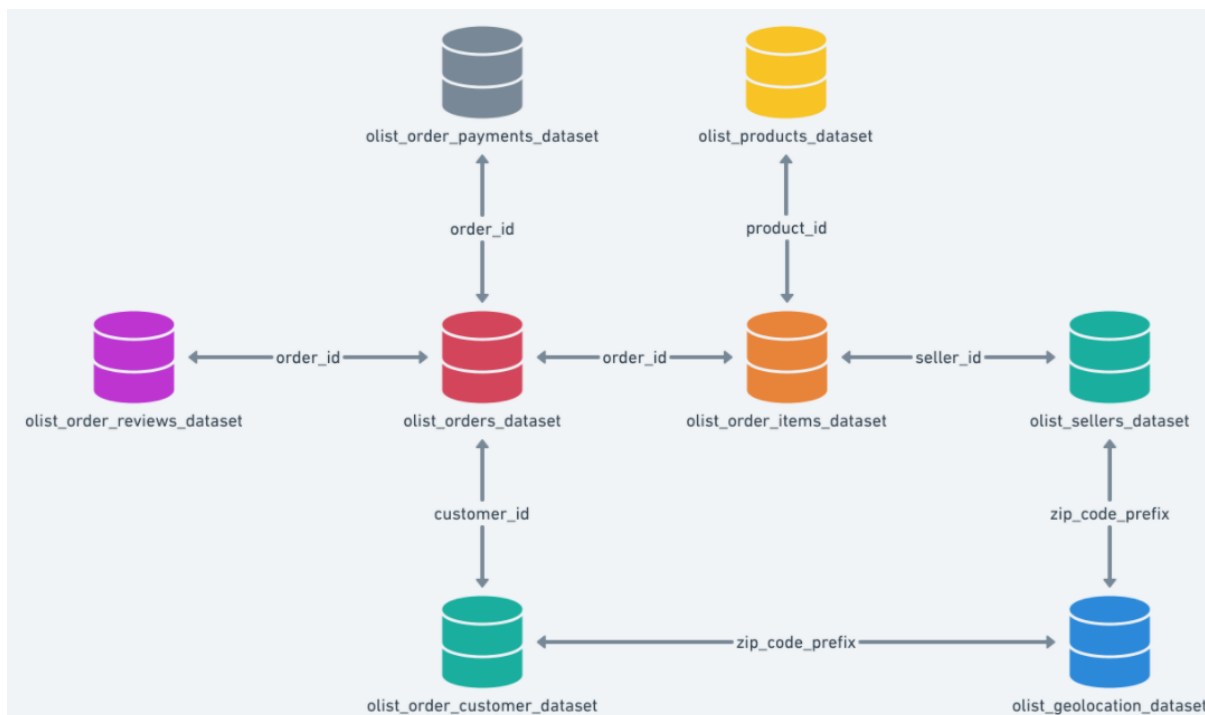


Fonte: elaborado pelo autor.

Essas empresas parceiras tentam cada vez mais fidelizar os vendedores que optam por anunciar com o nome da empresa, para isso, criam benefícios e programas de incentivo com o objetivo de aumentar as vendas dos vendedores e de ajudar vendedores que não vendem tanto a venderem mais.

Os dados que serão utilizados neste projeto estão divididos em 8 bases de dados que se interagem por variáveis conhecidas como variáveis chave, estas são variáveis que estão presentes em duas ou mais bases de forma idêntica, sendo assim, a principal função dessas variáveis é ajudar a unir as tabelas. No esquema abaixo vemos todas as bases da empresa, bem como, as variáveis chave.

Figura 2 - Esquema de organização das bases



Fonte: Kaggle (2022)

Como parte do lucro da empresa Olist vem diretamente da quantidade de produtos vendidos por vendedores cadastrados, a variável de interesse do estudo será definida como sendo: **os vendedores que não farão venda no mês seguinte**, assim seria possível executar alguma ação de incentivo para ajudá-los a realizar alguma venda.

Para melhor entendimento das variáveis que compõe este projeto, foi realizado um breve dicionário inicial para as variáveis das bases da empresa e apresentado neste texto como apêndice 1.

Desta forma, neste trabalho, deseja-se realizar um estudo sobre os vendedores da Olist para prever quais deles não vão vender no próximo mês. Este trabalho tem como objetivo principal, realizar o estudo por meio da aplicação de um modelo estatístico gerando um resultado claro e de fácil interpretação.

O modelo de Regressão Logística foi utilizado para prever se o vendedor efetuará alguma venda no próximo mês com os dados da empresa do varejo Olist.

Para atingir o objetivo desejado é necessário entender o objetivo do modelo através das reuniões com a área de negócio⁴. Posteriormente, criar a ABT⁵ (Analytical Base Table) a partir das informações obtidas nas reuniões com a área de negócio. Entender também, o comportamento das variáveis que compõe a ABT. Realizar a análise descritiva dos dados. Propor modelos para prever se o vendedor vai efetuar alguma venda no próximo mês. Entender quais fatores são mais influentes dado o objetivo do estudo.

A importância deste trabalho se deve ao fato que, com o repentino crescimento do e-commerce, seu domínio no mercado de varejo e esta nova estratégia de vendedores parceiros, monitorar os vendedores parceiros vem se tornando uma tarefa cada vez mais importante.

As empresas deste ramo utilizam diversas técnicas para chegar cada vez mais próximas desta informação. Com isso, saber se o vendedor vai efetuar a venda no mês seguinte ou não se tornou uma informação muito valiosa. Entre as várias técnicas que podem ser utilizadas para chegar a esta informação, o modelo de regressão logística pode ser adequado para esta situação, tendo em vista que a variável resposta é dicotômica.

Como mostrado por Uema e Lazzari (2008, pg.77) é notável o crescimento do número de pedidos no mercado virtual (e-commerce) a partir dos anos 2000 para a empresa estudada pelos autores (Submarino).

Com a Olist não foi diferente, desde a sua criação, em 2015, ela cresce consideravelmente a cada ano e durante a pandemia da COVID-19 seguiu crescendo:

“A pandemia acelerou processos de digitalização de comércios em todo o mundo, fazendo crescer o e-commerce de forma surpreendente. Nesse contexto, os recursos serão essenciais para que possamos continuar investindo no ecossistema e empoderando comerciantes em diferentes frentes” – Tiago Dalvi, founder e CEO do olist (Oliveira, 2021).

⁴ Área de negócio: este termo é utilizado nas empresas para se referir a área que está sendo consultada, por exemplo, ao fazer um estudo sobre o RH da empresa, a área de negócio será todo o departamento de RH e a pessoa de negócio será o funcionário com que se tem contato.

⁵ ABT: Analytical Base Table é o termo que faz referência a base de modelagem, ou seja, uma base de dados limpa e tratada, pronta para gerar um modelo.

2 EMPRESA OLIST

Neste capítulo será apresentada a empresa Olist, sua história e as principais características.

O Olist é uma startup que foi fundada em 2015 e é uma empresa do segmento de tecnologia que atua no setor de E-commerce, conhecido por ajudar e facilitar pequenos comércios. A empresa começou ajudando pequenos artesãos a venderem seus produtos em grandes e renomados marketplaces.

Segundo Fonseca (2022), o Olist é uma startup que conecta empreendedores a marketplaces, ou permite que eles criem lojas virtuais próprias. Hoje, atende 100 mil empreendedores em mais de 180 países.

Com missão de empoderar o comércio, o Olist vem surpreendendo e gerando resultados surpreendentes desde sua criação em 2015 e mesmo na Pandemia da COVID-19 surpreendeu gerando resultados incríveis. Adquiriu o título de “unicórnio” (grupo de startups avaliadas em mais de 1 bilhão) em 2021 quando foi avaliada em um valor superior a 1,2 bilhão de reais.

Para o Olist é importante que os vendedores parceiros efetuem vendas, assim a empresa gera maior lucro, já que os planos do Olist são associados a quantidade vendida. Dito isso, efetuar uma campanha para entender quais vendedores e principalmente, quais características dos que efetuarão venda no próximo mês é um possível caminho para selecionar estratégias que alavanquem as vendas dos vendedores mais ociosos.

3 REGRESSÃO LOGÍSTICA

Conhecida a partir de meados do século XX, a Regressão Logística é considerada uma das técnicas mais utilizadas para previsão de eventos. Embora não seja a técnica mais avançada e capaz de gerar o melhor resultado, a Regressão Logística é muito utilizada graças a seu poder de interpretabilidade dos parâmetros do modelo, ou seja, diferentemente de muitos outros modelos caixa preta⁶, ela tem interpretação.

A Regressão Logística é uma técnica estatística que pode ser utilizada para todos os tipos de dados, mas é comumente utilizada para dados categóricos, ou seja, dados qualitativos. Além disso, o objetivo da técnica é modelar a relação “logística” entre uma variável resposta dicotômica (0 ou 1) e várias variáveis explicativas categóricas e/ou numéricas (CABRAL, 2013).

Vale ressaltar que existem aplicações de modelos de Regressão Logística para previsão de eventos com mais de 2 classes, mas como o foco do presente trabalho é a previsão de um evento dicotômico este tópico não será abordado.

3.1 Modelo de Regressão Logística

Hosmer e Lemeshow (2013) mostram que a partir de um vetor de p variáveis independentes $\mathbf{x}^t = (x_1, x_2, \dots, x_p)$ com probabilidade condicional de que o evento de interesse seja $Y = 1$, denotada pela equação:

$$P(Y = 1|\mathbf{x}) = \pi(\mathbf{x}). \quad (3.1)$$

Assim, o logito do modelo de regressão logística múltipla será dado por:

$$\text{logit}(\mathbf{x}) = g(\mathbf{x}) = \ln\left(\frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})}\right) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p. \quad (3.2)$$

A equação da regressão logística múltipla é:

$$\pi(\mathbf{x}) = \frac{e^{g(\mathbf{x})}}{(1 + e^{g(\mathbf{x})})}. \quad (3.3)$$

⁶ Modelos caixa preta são modelos potentes capazes de gerar um resultado preciso, porém esses modelos não explicam como chegam no resultado (não são interpretáveis).

De tal forma que $g(x)$ é a equação linear

$$g(x) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p. \quad (3.4)$$

Os valores de $\beta_1, \beta_2, \dots, \beta_p$ são referentes à influência das variáveis $X_1, X_2, \dots, X_p$ na previsão.

3.2 Estimação dos parâmetros $\hat{\beta}_j$

Para estimar os valores de β_j pode ser utilizado o método da máxima verossimilhança, ou seja, maximiza-se a função de verossimilhança a fim de obter a melhor estimativa para o parâmetro β_j .

“O método da Máxima Verossimilhança consiste em estimar os parâmetros de um modelo utilizando as estimativas que tornam máximo o valor da função de verossimilhança. Isso é equivalente a encontrar o valor para o parâmetro que torna mínima a função de log-verossimilhança negativa.” – João L. F. Batista (Batista, 2009).

Seja β o vetor dos parâmetros β_j , seja $\pi(x_i)$ o valor de $\pi(x)$ para a observação i , seja y o vetor da variável de interesse e seja n o número de observações na amostra, então a equação a ser maximizada é:

$$L(\beta|y) = \prod_{i=1}^n \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i}. \quad (3.5)$$

Para facilitar, aplica-se o logaritmo na equação (3.5), o resultado é:

$$l(\beta|y) = \ln(L(\beta|y)) = \sum_{i=1}^n \{y_i \ln[\pi(x_i)] + (1 - y_i)[1 - \pi(x_i)]\}. \quad (3.6)$$

Assim, para encontrar o vetor β que minimiza a equação (3.6) aplica-se a derivada parcial em $l(\beta|y)$. Para isso deriva-se $l(\beta|y)$ em relação a β_j para $j = 0, 1, 2, \dots, p$. No qual, p é o número de parâmetros β , obtendo-se então a seguinte equação:

$$\frac{\partial L(\beta|y)}{\partial \beta_j} = \begin{cases} \sum_{i=1}^n [y_i - \pi(x_i)], & \text{para } j = 0 \\ \sum_{i=1}^n x_{i,j} [y_i - \pi(x_{i,j})], & \text{para } j > 0 \end{cases} \quad (3.7)$$

Ao igualar a equação (3.7) a zero, obtém-se as equações de estimação para o $\hat{\beta}_j$.

3.3 Teste da razão de Verossimilhança e Teste Z-Wald

Após efetuar o ajuste de um modelo sempre precisa-se verificar se todos os parâmetros ajustados são significativos, para isso existem testes estatísticos, veremos dois deles, o teste da razão de Verossimilhança e o teste Z-Wald.

O teste da razão de Verossimilhança é o teste global da regressão logística, ou seja, esse teste vai verificar se pelo menos um parâmetro utilizado é significativo e o teste é feito a partir da razão do log da verossimilhança sem as variáveis de interesse (L_s) pela verossimilhança com as variáveis de interesse (L_c), gerando assim a estatística do teste da seguinte maneira:

$$G = -2 \ln \left[\frac{L_s}{L_c} \right]. \quad (3.8)$$

Ou também,

$$G = -2[\ln(L_s) - \ln(L_c)]. \quad (3.9)$$

Seja θ um vetor com todas as p variáveis do modelo e $\mathbf{0}$ um vetor nulo com p elementos, então as hipóteses do teste são dadas por:

$$\begin{cases} H_0: \theta = \mathbf{0} \\ H_a: \theta \neq \mathbf{0} \end{cases} \quad (3.10)$$

Como G é um valor de χ^2 com o número de variáveis (β_j) graus de liberdade, então a regra de rejeição é, se o valor de G for maior que o valor do χ^2 tabelado com o nível de significância desejado, então rejeita-se H_0 e conclui-se que pelo menos um β_j é diferente de zero com o nível desejado de confiança $(1 - \alpha)$.

O teste Z-Wald, por sua vez, é o teste individual de cada parâmetro estimado para equação, ou seja, esse teste verifica se cada parâmetro β_j é diferente de zero individualmente. Hosmer e Lemeshow (2013) definem estatística do teste da seguinte maneira:

$$W_j = \frac{\hat{\beta}_j}{\widehat{SE}(\hat{\beta}_j)} \sim N(0,1). \quad (3.11)$$

No qual, $\widehat{SE}(\hat{\beta}_j)$ é o desvio padrão de $\hat{\beta}_j$:

$$\widehat{SE}(\hat{\beta}_j) = \sqrt{\widehat{Var}(\hat{\beta}_j)}. \quad (3.12)$$

Como W_j segue uma distribuição normal padrão, então W_j^2 seguirá uma distribuição χ_1^2 com um grau de liberdade. As hipóteses do teste são as seguintes:

$$\begin{cases} H_0: \hat{\beta}_j = 0 \\ H_a: \hat{\beta}_j \neq 0 \end{cases} \quad (3.13)$$

A regra de rejeição é, se χ^2 calculado $>$ χ^2 tabelado, então rejeita-se H_0 e o parâmetro é significativo com $(1 - \alpha)$ de confiança estatística.

3.4 Intervalo de confiança dos Parâmetros

Geralmente na estatística escuta-se muito a frase: um número não é só um número, um número é um intervalo. Seguindo nesse pensamento, os parâmetros estimados não são somente o valor estimado, portanto precisa-se estimar o intervalo para cada parâmetro do modelo.

O intervalo de confiança do parâmetro é baseado no teste de Wald, e Hosmer e Lemeshow o definem da seguinte maneira:

$$IC\ 100(1 - \alpha)\% = [\hat{\beta}_j \pm z_{1-\alpha/2} \widehat{SE}(\hat{\beta}_j)]. \quad (3.14)$$

Para $j = 0, 1, 2, \dots, p$.

3.5 Deviance residual

Existem diversas formas de avaliar a qualidade do ajuste, uma delas é o Deviance residual. Segundo Hosmer e Lemeshow (2013) para um valor ajustado $\hat{y}_j$ definido como

$$\hat{y}_j = m_j \hat{\pi}_j = m_j \left\{ \frac{e^{\hat{g}(x_j)}}{1 + e^{\hat{g}(x_j)}} \right\}. \quad (3.15)$$

o Deviance residual é definido por:

$$d(y_j, \hat{\pi}_j) = \pm \left[2 \left[y_j \ln \left(\frac{y_j}{m_j \hat{\pi}_j} \right) + (m_j - y_j) \ln \left(\frac{(m_j - y_j)}{m_j (1 - \hat{\pi}_j)} \right) \right] \right]^{1/2}. \quad (3.16)$$

Assim, o sinal $+$ ou $-$ é o mesmo sinal de $(y_j - m_j \hat{\pi}_j)$. Quando $y_j = 0$ ou $y_j = m_j$ a fórmula do deviance residual será reduzida, ficando então das seguintes formas:

- $y_j = 0$

$$d(y_j, \hat{\pi}_j) = -\sqrt{2m_j |\ln(1 - \hat{\pi}_j)|}; \quad (3.17)$$

- $y_j = m_j$

$$d(y_j, \hat{\pi}_j) = -\sqrt{2m_j |\ln(\hat{\pi}_j)|}. \quad (3.18)$$

Assim, a estatística (D) baseada no deviance residual é definida como:

$$D = \sum_{j=1}^N d(y_j, \hat{\pi}_j). \quad (3.19)$$

No qual, a estatística D seguirá uma χ^2 com $N - (p + 1)$. Em que p é o número de parâmetros β sem contar o intercepto.

3.6 Odds Ratio

A interpretação do modelo de regressão logística, diferentemente do modelo de regressão linear, não é feita de forma direta. Para interpretar os parâmetros do modelo de regressão logística precisa-se utilizar a razão das chances (Odds Ratio).

A odds ratio ajuda com a interpretação do parâmetro explicando a relação da variável em questão com a probabilidade de sucesso. Por exemplo, se a variável sexo foi inserida no modelo como β_1 e está da forma: feminino = 1 e masculino = 0, então a interpretação da sua odds ratio ficará: a chance do sexo feminino cometer o evento de interesse é $|e^{\beta_1} - 1|$ vezes maior (ou menor, dependendo do valor da expressão e^{β_1} , se for maior que 1, então a interpretação será maior, caso contrário a interpretação se dá por menor) que a chance do sexo masculino cometer o evento de interesse.

Outra forma possível para esta interpretação seria: a chance do sexo feminino cometer o evento de interesse é e^{β_1} vezes a chance do sexo masculino cometer o evento de interesse.

Para ilustrar o que foi explicado segue o exemplo: Supondo que o modelo esteja interessado em prever se o animal é agressivo ou não (agressivo = 1 e não agressivo = 0) e supondo ainda que a variável tipo do animal (gato = 1 e cachorro = 0) tenha o parâmetro $\beta = 1$, então para interpretar a odds ratio (razão das chances), faz-se $e^1 \cong 2,71$. A partir deste valor podemos interpretá-la da seguinte forma: A chance do

animal ser agressivo quando o tipo dele é gato será 1,71 (171%) vezes maior do que quando o tipo do animal é cachorro.

A interpretação para variáveis numéricas (quantitativas) é diferente das variáveis categóricas (qualitativas), segue o exemplo para explicação da interpretação de variáveis numéricas: Supondo o exemplo acima, porém, no lugar da variável tipo do animal, tem-se a variável idade em anos, supondo ainda que o parâmetro $\beta = 1$, a interpretação da odds ratio ($e^1 \cong 2,71$) é: A cada incremento de 1 ano na idade do animal, a chance de ele ser agressivo aumenta em 1,71 (171%).

Assim, a equação para a odds ratio do parâmetro β_j é:

$$ODDs(\beta_j) = e^{\beta_j}. \quad (3.20)$$

Com $j = 1, 2, \dots, p$.

4 MÉTRICAS DE QUALIDADE DE AJUSTE

4.1 Objetivo do modelo

Silva (2019) explica que um modelo sempre contém um desses 3 objetivos: Decisão, Ordenação e Estimação. Comumente nas empresas o objetivo do modelo é decidido através de uma reunião com a área de negócio onde são feitas perguntas como: “Para que irá servir este modelo?”, “Como usarão este modelo?” e “Qual é o intuito do modelo?”. Essas perguntas serão feitas com foco em extrair o máximo de informação útil para os analistas.

Os modelos de decisão são utilizados quando não há restrições, ou seja, estimamos uma probabilidade para cada indivíduo e definimos um ponto de corte (cut-off), então todas as observações com probabilidade maior que o ponto de corte são classificadas como 1 (evento de interesse) e todos com probabilidade menor que o ponto de corte são classificados como 0 (evento de não interesse).

Os modelos de ordenação são usados quando há uma restrição (normalmente financeira) e por conta disso, é necessário ordenar as probabilidades da maior até a menor. Um exemplo que ilustra os modelos de ordenação é um modelo que prevê os cinco clientes mais prováveis de cometer fraude. É necessário ordenar para identificar quais serão as cinco observações com maior probabilidade a fraudar.

Por outro lado, nos modelos de estimação é necessário utilizar todas as probabilidades geradas pelo modelo individualmente para entender qual ação deve ser tomada, para isso a estimativa deve ser a mais próxima da realidade. Normalmente utiliza-se este modelo juntamente com outra informação para multiplicar pela probabilidade estimada, gerando um valor esperado como resultado.

Vale ressaltar que em todos os casos busca-se ajustar o melhor modelo, a diferença está em quais métricas serão utilizadas para avaliar cada modelo com base no seu objetivo.

4.2 Matriz de confusão

A matriz de confusão é gerada com os dados reais e os valores previstos pelo modelo. A partir dela é possível obter diversas medidas de qualidade de ajuste, como: Acurácia, Precisão, Recall, Sensibilidade, Especificidade, Erro, etc.

Porém a matriz de confusão é mais indicada para modelos com objetivo de decisão, afinal um ponto de corte foi definido para a probabilidade estimada do

modelo. Além disso, as métricas que serão abordadas com mais detalhes neste estudo são: Acurácia, Precisão e Recall.

Silva (2021) apresenta a matriz de confusão como a imagem abaixo:

Figura 3 – Matriz de confusão

		MODELO	
		0	1
REAL	0	True Negative TN	False Positive FP
	1	False Negative FN	True Positive TP

Fonte: Adaptação de Fawcett (2006).

A partir da matriz de confusão para calcular as métricas faz-se da seguinte maneira:

$$Acurácia = \frac{TN + TP}{TN + FP + FN + TP}. \quad (4.1)$$

A interpretação da acurácia é: Taxa de acerto geral do modelo.

$$Precisão = \frac{TP}{FP + TP}. \quad (4.2)$$

A interpretação da precisão é: De todos que o modelo previu como 1 (evento de interesse), quantos realmente são 1. Ou seja, a porcentagem de eventos de interesse que o modelo acertou.

$$Recall = \frac{TP}{FN + TP}. \quad (4.3)$$

A interpretação do recall é: De todos que são 1, quantos o meu modelo está contemplando. Ou seja, a porcentagem de eventos de interesse capturada pelo modelo.

Vale ressaltar que a precisão e o recall têm relação inversamente proporcional, ou seja, ao mudar o ponto de corte se o recall aumenta a precisão diminui e o contrário também é verdade.

4.3 Lift

O Lift é uma métrica de qualidade de ajuste recomendada principalmente para modelos de ordenação, é uma métrica muito poderosa para convencer a área de negócio, pois mostra diretamente quantas vezes a ordenação do modelo é melhor do que a escolha aleatória.

O cálculo é bem simples, dada uma restrição dos primeiros $X\%$ dos dados, o cálculo do lift será a taxa média dos primeiros $X\%$ dividido pela taxa média da base toda. Ou seja, faz-se a média das probabilidades estimados pelo modelo e a média dos primeiros $X\%$ valores ordenados e se divide o valor dos $X\%$ pelo valor total.

Seja a taxa dos primeiros $X\%$ denominada por T_X e a taxa da base toda denominada por T_{100} , o lift será:

$$Lift = \frac{T_X}{T_{100}}. \quad (4.4)$$

4.4 Métricas associadas ao erro

As métricas associadas ao erro são normalmente recomendadas para modelos com objetivo de estimação, já que é muito importante que a estimativa seja a mais precisa possível em todas as observações.

Após a execução do modelo podemos calcular valores como a soma dos quadrados da regressão (SSR) e a soma dos quadrados do erro (SSE) que juntamente com a soma dos quadrados totais (SST) geram uma série de métricas de qualidade de ajuste, das quais duas foram selecionadas para este projeto.

Sejam $\hat{y}_i$ os N valores preditos pelo modelo e y_i os N valores originais, seja ainda $\bar{y}$ a média dos valores originais. A SST é definida como:

$$SST = \sum_{i=1}^N (y_i - \bar{y})^2. \quad (4.5)$$

Já a SSR é definida como:

$$SSR = \sum_{i=1}^N (\hat{y}_i - \bar{y})^2. \quad (4.6)$$

E a SSE é:

$$SSE = \sum_{i=1}^N (\hat{y}_i - y_i)^2. \quad (4.7)$$

Uma das principais métricas para um modelo de estimativa é o RMSE, ou seja, o erro quadrático médio do modelo. O RMSE é definido pela seguinte equação:

$$RMSE = \sqrt{\frac{SSE}{N}} = \sqrt{\frac{\sum_{i=1}^N (\hat{y}_i - y_i)^2}{N}}. \quad (4.8)$$

O critério de informação de Akaike (AIC) é outra métrica muito conhecida e utilizada para modelos de estimação, sendo k o número de parâmetros no modelo (incluindo o intercepto β_0), esse critério é definido como:

$$AIC = N \ln \left(\frac{SSE}{N} \right) + 2k. \quad (4.9)$$

Para as duas últimas métricas citadas (RMSE e AIC) é muito importante minimizá-las, pois as métricas são associadas ao erro e o foco é minimizar o erro.

5 TRATAMENTO DOS DADOS

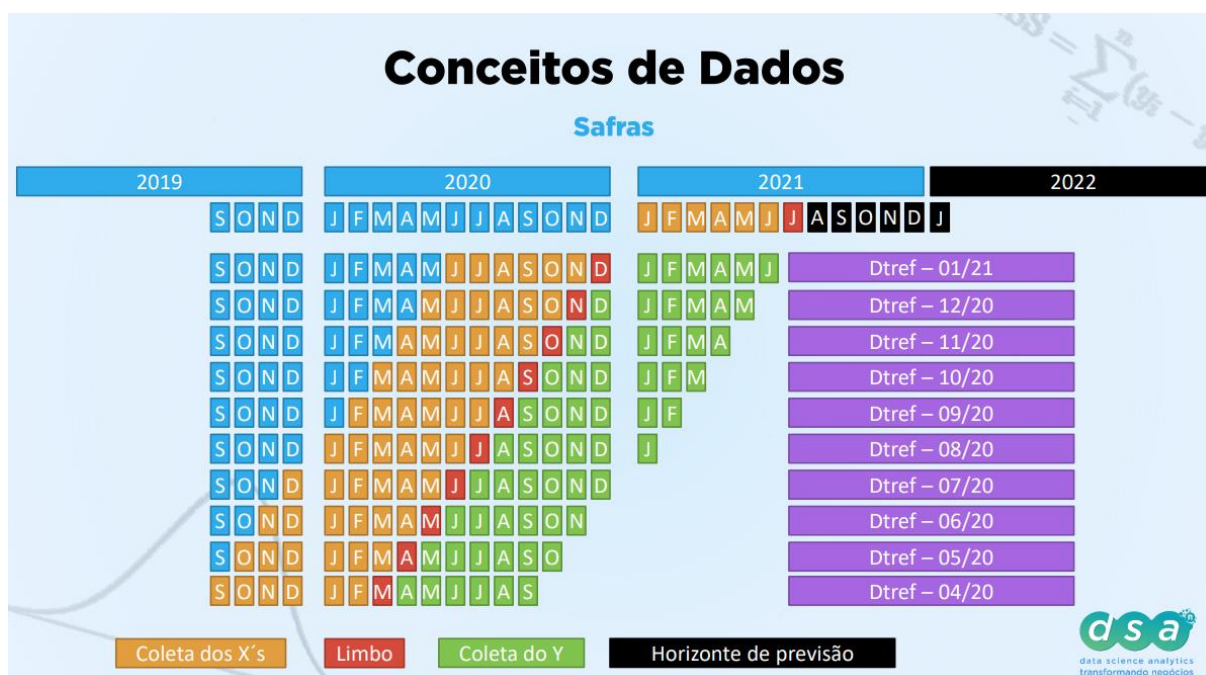
5.1 Safras

Comumente utilizadas em áreas de modelagem, as safras são capazes de sanar dois problemas de modelagem, a falta de dados e a sazonalidade de modelos que performam bem apenas em determinada parte do ano. Ao utilizar safras nos dados, o volume de dados aumenta, ou seja, o modelo será capaz de aprender com um maior conjunto de dados, o que provavelmente fará este modelo desempenhar melhor na previsão de informações futuras, pois ele será mais generalista para situações assim, evitando criar um modelo extremamente específico que não desempenha bem com novos conjuntos de dados.

Outro ponto em que as safras ajudam é na sazonalidade do modelo, ao aplicar safras elimina-se a sazonalidade dos dados diminuindo a probabilidade de gerar um modelo que performa bem em determinada parte do ano, mas não performa bem em outra.

Para utilizar safras nos dados é necessário definir o horizonte de previsão, o nível das safras e a data limite de aplicação das safras. Para ilustrar e exemplificar o uso de safras segue a imagem retirada do material de Silva (2021).

Figura 4 – Aplicação de safras nos dados.



Fonte: Materiais ASN.Rocks (2021)

Na figura acima, as letras são referentes aos meses, as linhas “Dref” que estão na cor roxa mostram as datas de referência do primeiro mês em que serão coletados

os Y para criar o modelo. O período de coleta das variáveis X's (variáveis preditoras) é exibido em laranja, enquanto isso, o período de coleta da variável y (variável resposta) está em verde.

Nota-se que o horizonte de previsão está bem definido (agosto de 2021 até janeiro de 2022), o nível das safras está mensal e a data limite é setembro de 2019.

A partir dessa divisão o próximo passo é empilhar os dados⁷, ou seja, os dados em laranja são empilhados e os dados em verde também. Assim, o resultado vai ser uma tabela com um volume de dados maior e com seis meses de dados para gerar um modelo que preverá os seis meses seguintes ao limbo. Esses seis meses serão utilizados para validação do modelo.

O que foi chamado de limbo na Figura 4 é o período em que não se pode efetuar a coleta dos dados nem para modelagem nem para validação do modelo. Para facilitar o entendimento segue um exemplo: O gerente precisa saber quais funcionários pedirão demissão em fevereiro, mas ele precisa dessa informação um mês antes (dia 1 de janeiro). No exemplo anterior pode-se utilizar os dados até dezembro para prever quem pedirá demissão em fevereiro, mas não se pode utilizar os dados de janeiro, pois a resposta deve ser dada antes no primeiro dia do mês de janeiro.

5.2 Dados faltantes

O modelo de regressão logística é um modelo que não admite dados faltantes, assim existem algumas soluções possíveis para essa situação. Uma estratégia inicial é remover todas as observações com pelo menos uma variável com dados faltantes, porém, com esta estratégia perde-se a observação e isso pode acarretar um modelo pior caso o número de dados faltantes seja alto.

Outro caminho que pode ser seguido é preencher os dados faltantes com alguma estimativa. E essa estimativa pode ser uma média ou mediana no caso de dados numéricos ou uma moda para dados categóricos. Mas essas não são as únicas estratégias, pode-se também gerar outros modelos para preencher os dados faltantes, por exemplo, criar uma regressão linear para estimar o valor da variável numérica é uma solução plausível.

Cabe ao analista escolher uma ou mais estratégias para sanar o problema de dados faltantes, mais de uma estratégia pode ser usada no mesmo conjunto de dados.

⁷ Empilhar os dados é o ato de pegar bases de dados e juntá-las gerando assim uma base de dados onde o número de linhas será a soma das linhas de todas bases de dados juntas.

Por exemplo, se a variável X tem muitos outliers (pontos extremos) pode-se utilizar a mediana para os dados faltantes, já se a variável Z não tem muitos outliers pode-se utilizar a média para os valores faltantes.

6 ADAPTAÇÃO DA METODOLOGIA CRISP-DM

A metodologia deste trabalho tem início com uma atualização da metodologia CRISP-DM⁸ utilizada pela empresa Simplilearn (a qual pode ser visualizada no link: <https://youtu.be/X3paOmcrtjQ>) que divide o processo de ciência de dados em 7 partes, sendo elas:

1. *Business problem* (Problema de negócios);

Nesta etapa o objetivo é questionar a área de negócio e entender três principais dúvidas: O que é o problema? Para que será usado? E como essa solução será usada? Além disso é de extrema importância entender o contexto e o cenário da empresa, pois para algumas empresas o mesmo termo (com por exemplo, “churn”) pode ter definições diferentes.

2. *Data acquisition* (Aquisição dos dados);

Nesta etapa o objetivo é coletar os dados. Esta etapa normalmente é feita buscando os dados na internet, por meio de API's (*Application Programming Interface*), repositórios online, etc. Porém, nada impede que nesta etapa seja feita uma amostragem e a partir disso uma aplicação de questionários para a coleta dos dados a serem estudados.

3. *Data preparation* (Preparação dos dados);

Nesta etapa o objetivo é limpar e transformar os dados para um formato possível de analisar. Por conta disso, o dado segue por um processo denominado ETL (*Extract, transform, load*) no qual o dado é extraído do local de origem, transformado (tratando inconsistências, traduzindo valores codificados, agrupando dados, selecionando apenas colunas úteis para o problema, resumindo variáveis, criando variáveis, etc.) e finalmente carregado. Segundo a empresa SAS⁹:

“Nesse processo, os dados são retirados (extraídos) de um sistema-fonte, convertidos (transformados) em um formato que possa ser analisado, e armazenados (carregados) em um armazém ou outro sistema. Extração, carregamento, transformação (ELT) é uma abordagem alternativa, embora relacionada, projetada para jogar o processamento para o banco de dados, de modo a aprimorar a performance”. (SAS)

⁸ CRISP-DM (*Cross Industry Standard Process for Data Mining*) ou Processo Padrão Inter-Indústrias para Mineração de Dados que surgiu em 1996 como forma de apoio ao processo de descoberta do conhecimento

⁹ SAS é uma empresa privada que possui um dos principais softwares de análises de dados.

Além de fazer os processos anteriores, nesta etapa é criada a ABT (*Analytical Base Table*), base de dados que é utilizada para construir o modelo) em que serão feitas as análises. Para criar a ABT é necessário muito entendimento de negócio, contato com a área de negócio para sanar eventuais dúvidas e muita criatividade.

4. *Exploratory data analysis* (Análise exploratória dos dados);

Nesta etapa o objetivo é analisar, entender e aprofundar nos dados. Entender o comportamento das variáveis em questão facilita o processo de modelagem futuramente, além de que, em muitos casos a análise exploratória (quando bem feita) pode solucionar muitos problemas.

5. *Data modeling* (Modelagem);

Nesta etapa o objetivo é criar modelos com diversas técnicas diferentes, validá-los, escolher o melhor modelo a partir da métrica de qualidade de ajuste escolhida e treiná-los em uma “base virgem” (conhecida também por base de teste), ou seja, uma base que não foi utilizada anteriormente, podendo até ser uma base recém coletada.

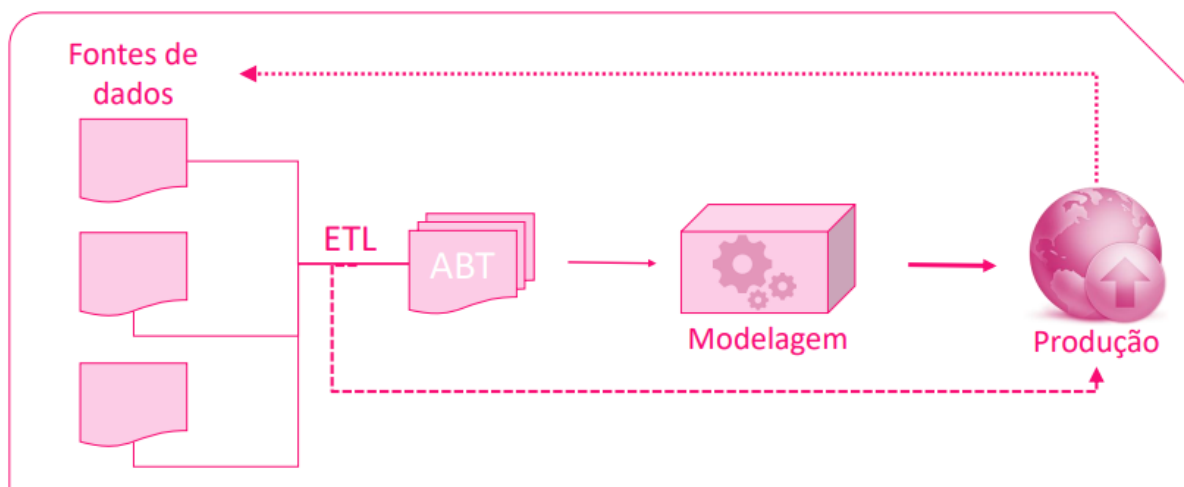
6. *Visualization and communication* (Visualização e comunicação);

Nesta etapa o objetivo é mostrar os resultados obtidos de forma clara e em uma linguagem que todos possam entender o que foi feito. Isto é um desafio em muitos casos pois explicar um problema complexo de forma simples nem sempre é fácil.

7. *Deploy and maintains* (Implementação e manutenção);

Nesta etapa o objetivo é fazer o modelo funcionar por conta própria, ou seja, os dados vão entrar na base de dados (seja por uma coleta de forma ativa ou passiva) e passarão pelo modelo, onde serão analisados gerando a resposta do modelo. Para esse processo funcionar continuamente é necessário que o fluxo por onde o dado passará seja todo automatizado, como na figura abaixo. Além disso, este processo necessita ser monitorado para garantir que o modelo utilizado está desempenhando com o esperado.

Figura 5 - Processo de modelagem automatizado.



Fonte: Materiais ASN.Rocks (2021)

Porém para este projeto algumas alterações serão feitas em alguns passos, sendo elas:

- Parte 1: todo contato com a área de negócio será feito através de reuniões com a CEO Adriana M. M. Silva, em que ela irá simular a gestora da empresa Olist cujos dados foram publicados na plataforma Kaggle.¹⁰
- Parte 2: os dados serão adquiridos através da plataforma Kaggle, no caso da empresa de varejo Olist².
- Parte 3: os dados serão preparados, transformados e tratados com a utilização de, pelo menos, um dos softwares a seguir: SQL, R e Python. Podendo então, haver mais de um software para esta e para todas as etapas que necessitem do uso de softwares. Durante a criação da ABT, caso haja necessidade, reuniões com a área de negócio poderão acontecer, todas com o objetivo de entender o contexto e captar ideias para a criação de variáveis.
- Parte 4: esta etapa será feita por, pelo menos, um dos seguintes softwares: R, Python e SAS. Caso necessário, mais de um software será usado.
- Parte 5: a métrica para decisão do melhor modelo será decidida após a reunião da parte 1 (reunião de negócio com objetivo de entender o problema), pois

¹⁰ Kaggle é um site que contém diversas bases de dados e projetos, com diversos focos diferentes, nele é possível encontrar alguns dados públicos da empresa Olist.

apenas após esta reunião será possível descobrir o objetivo do modelo. Para Silva (2019, 2020 e 2021) dentre os 3 objetivos existentes (Ordenação, Estimação e Decisão) algumas métricas se destacam para cada tipo de modelo. Caso o modelo tenha como objetivo ordenar, a métrica de qualidade de ajuste utilizada será o Lift. Por outro lado, se o modelo tiver como objetivo decidir, as possíveis métricas de qualidade de ajuste utilizadas serão: precision e recall. E finalmente se o modelo for um modelo de estimação, então as métricas de qualidade de ajuste utilizada serão: RMSE e AIC.

- Parte 6: esta etapa será feita ao longo de todo o projeto, a ideia é manter a linguagem do projeto a mais acessível possível e assim motivar e incentivar os alunos dos primeiros anos do curso de Estatística.
- Parte 7: colocar este modelo em produção¹¹ não faz sentido neste projeto, para colocá-lo em produção seria necessária a entrada contínua de dados, além disso, a manutenção do modelo também não é possível de ser feita, já que o modelo não entrará em produção.

¹¹ Colocar um modelo em produção é quando todo o processo de entrada de novos dados, limpeza, transformações, análise e geração de resultados é automatizado de certa forma que ao ocorrer uma nova entrada de dados, o resultado será gerado automaticamente.

7 DESENVOLVIMENTO

Inicialmente fez-se necessária uma reunião com a área de negócio para melhor entendimento do problema, bem como, para sanar algumas dúvidas iniciais observadas nas bases de dados obtidas no Kaggle.

7.1 Reunião com a área de negócio

Com foco em sanar as primeiras quatro perguntas obtidas: “O que é para a Olist um vendedor não efetuar uma venda?”, “Por que precisa da informação que será obtida?”, “Qual é o período de previsão?” e “A Olist possui recursos ilimitados para esse projeto? Se a resposta for não, qual é o valor disponível para o projeto?”. Assim, foi marcada a primeira reunião que pode ser visualizada através do link: https://drive.google.com/file/d/1zXNzzvNOAB3z2DYMlrDRhAI_aKyPIR8D/view?usp=sharing.

Através da reunião foi possível extrair algumas informações importantes para o andamento do projeto. Entre essas informações é interessante definir que para este projeto o foco é entender se o vendedor vai ter algum pedido no mês seguinte, ou seja, o evento observado é o pedido, assim, se o cliente fizer um pedido e depois efetuar o cancelamento do pedido ou houver algum problema com a entrega, isso não irá importar nesse projeto pois o pedido já foi feito.

Outro ponto importante é saber para que essa informação será usada, assim é mais fácil direcionar as apresentações e perguntas visando ajudar da melhor forma o cliente. Nesse caso, ao saber quais vendedores não receberão pedidos no próximo mês o objetivo da Olist é criar uma campanha de divulgação para o principal produto desses vendedores.

O conhecimento sobre o período de previsão é extremamente importante, essa informação dá suporte para a decisão de como estruturar a ABT de modelagem, bem como, saber se o uso de safras é necessário ou não. Na reunião foi obtida a informação que o período de previsão é o mês seguinte, ou seja, no último dia do mês é necessário prever quais vendedores não terão pedidos no mês seguinte inteiro.

Vimos que as métricas de qualidade de ajuste são recomendadas conforme o objetivo do modelo, portanto, entender esse objetivo é extremamente importante. Haja vista que os diferentes objetivos de modelo têm algumas características que os

diferenciam, então, é possível descobrir qual é o objetivo do modelo. Como existe uma limitação financeira, mesmo que desconhecida inicialmente, então, o objetivo de ordenação parece ser mais atraente para este projeto.

7.2 Criação da ABT

A criação da ABT fez-se através do software SQLiteStudio que trabalha com a linguagem SQL. O código foi uma adaptação do código criado pelo professor Sergio Zaccarelli disponibilizado como material de apoio no curso Data Science Analytics: Transformando Negócios. Após criada a ABT notou-se que é uma base de dados que contém 202 variáveis e 25462 linhas. O dicionário da ABT encontra-se no Apêndice 2.

7.3 Tratamento dos dados

Os dados obtidos através do SQL foram levados para o software R, onde todo o processo de tratamento, análise e modelagem foi feito.

Inicialmente foram deletadas variáveis com mais de 8000 dados faltantes, pois considerou-se desnecessário estimar um valor para uma variável em que mais de 30% dos valores são dados faltantes, assim, restaram 136 variáveis. Também foram deletadas linhas (vendedores) com 18 ou mais dados faltantes, ou seja, linhas em que mais de 18 variáveis eram dados faltantes, pois considerou-se prejudicial manter um indivíduo que possui mais de 13% das informações como dados faltantes, assim, restaram 19468 linhas (vendedores).

Após o tratamento das inconsistências, foi realizado o preenchimento dos dados faltantes (ou imputação¹²), onde o valor para tal preenchimento foi obtido através da análise descritiva da variável e do entendimento do processo de criação da ABT no SQL, isso pois, variáveis que seguiam distribuição Normal foram preenchidas com a média, já as variáveis que seguiam distribuição exponencial foram preenchidas com a mediana e outras variáveis, por exemplo, variáveis que contam os caracteres na descrição, portanto, se não há nenhuma descrição, é esperado que o resultado da variável seja zero.

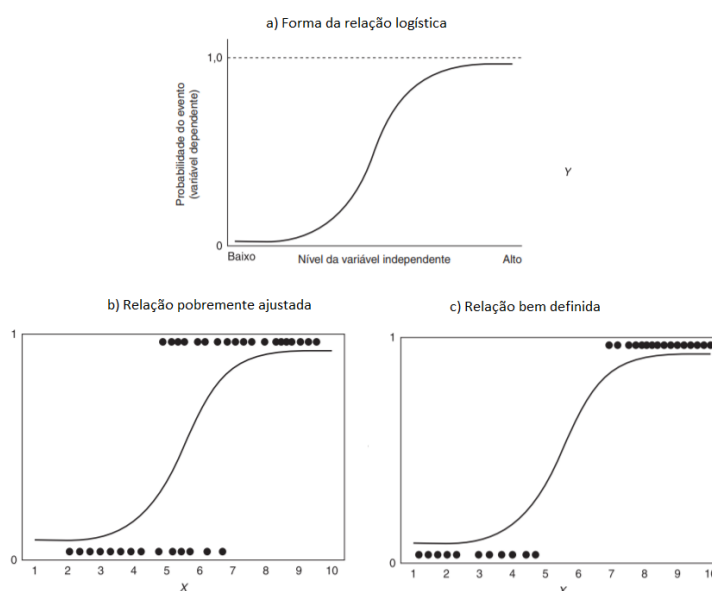
¹² Imputação é o termo utilizado no mercado de trabalho para se referir ao ato de preencher dados faltantes.

7.4 Análise descritiva

Nesta etapa foram feitas as análises univariadas e bivariadas de todas as variáveis, a partir disso, foram selecionadas as variáveis que apresentavam melhor resultado no modelo, outro critério foi manter duas das variáveis defasadas no tempo, optou-se por manter as pontas, por exemplo, a variável número de produtos no último mês e número de produtos nos últimos seis meses foram mantidas, enquanto número de produtos nos últimos dois, três, quatro e cinco meses foram retiradas.

O foco da análise descritiva foi principalmente entender se há diferença no comportamento da variável quando há a venda e quando não há venda. Para isso o objetivo com os gráficos de dispersão era encontrar um gráfico parecido com a figura a seguir:

Figura 6 – Objetivo do uso do gráfico de dispersão

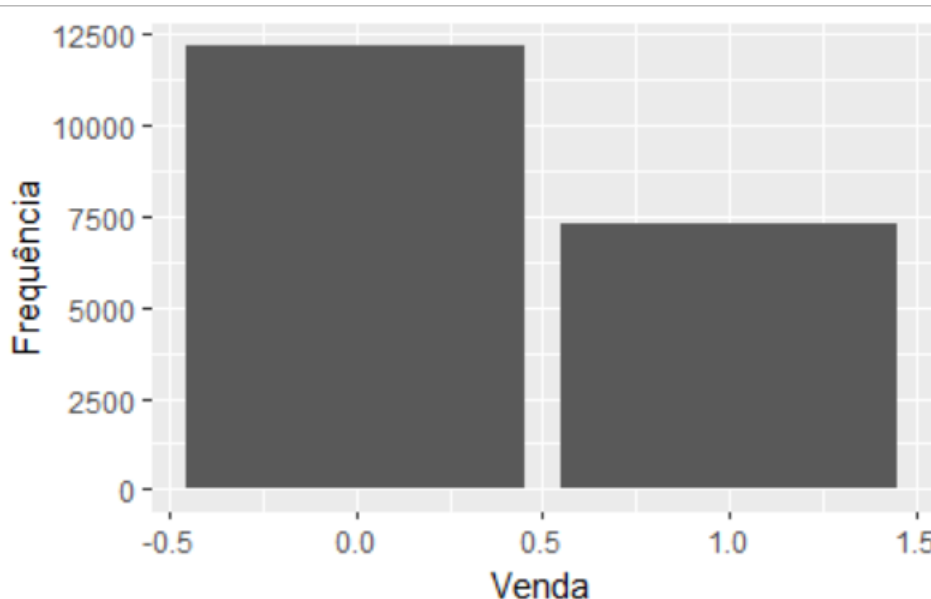


Fonte: Adaptado de Hair et al. (2009)

Outro gráfico analisado foi o gráfico boxplot dividido pela variável venda, o objetivo é entender se as variáveis se comportam de forma diferente quando há e quando não há venda.

A variável venda que é a variável de estudo se comporta de forma que o valor 1 representa os vendedores que não venderam e o valor 0 representa os vendedores que venderam. A proporção de vendedores que não fizeram venda é de 0,375077 o que corresponde a 7302 indivíduos que não fizeram venda. No Gráfico 1 é apresentado estes resultados.

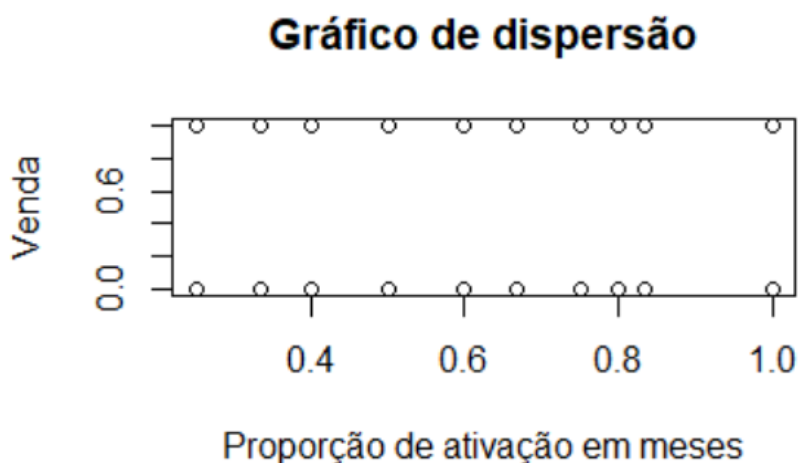
Gráfico 1 – Gráfico de barras da variável fl_venda



Fonte: Elaborado pelo autor.

A variável proporção de ativação em meses é proporção de ativação do vendedor em meses, ou seja, se um vendedor teve venda nos últimos 3 meses e está ativo nos 3 meses, e foi indicado como 1. Se tem venda nos últimos 2 meses, mas entrou nos últimos 2 meses, foi indicado como 1. Se teve venda em 2 meses, mas está ativo nos 3 meses, então 0,66. Esta variável não contém valores faltantes. No Gráfico 2, é apresentado o gráfico de dispersão com a variável resposta. Pode se observar que a ocorrência de valores 0 (zero) ou 1 (um) não está relacionado a proporção de ativação.

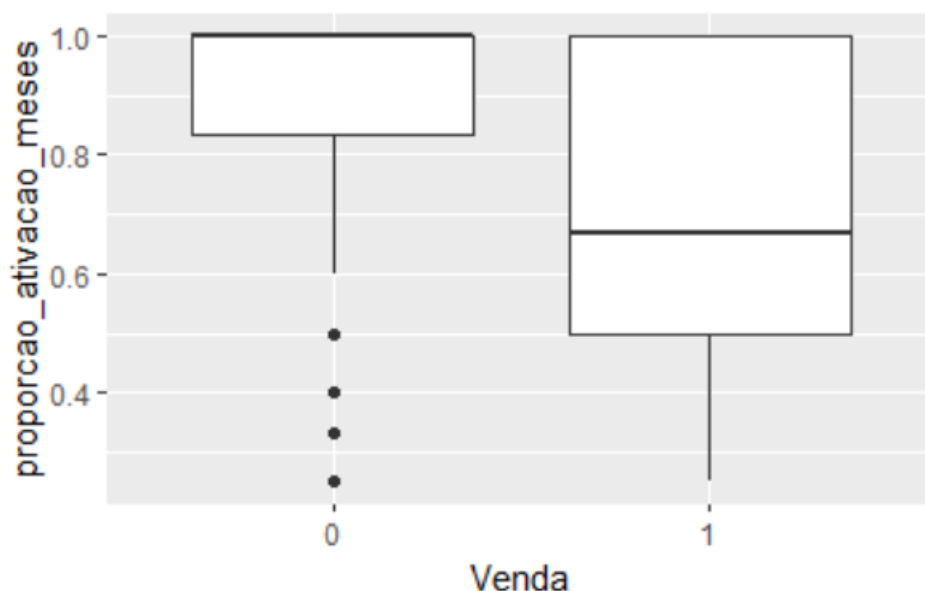
Gráfico 2 – Gráfico de dispersão das variáveis proporção ativação meses e venda



Fonte: Elaborado pelo autor.

Observa-se que a análise visual não gera conclusões claras a partir deste gráfico, o que significa que talvez não seja uma variável boa para o modelo. No Gráfico 3, é apresentado o Boxplot da proporção de ativação entre os vendedores com venda e sem venda.

Gráfico 3 – Boxplot da proporção ativação meses por venda

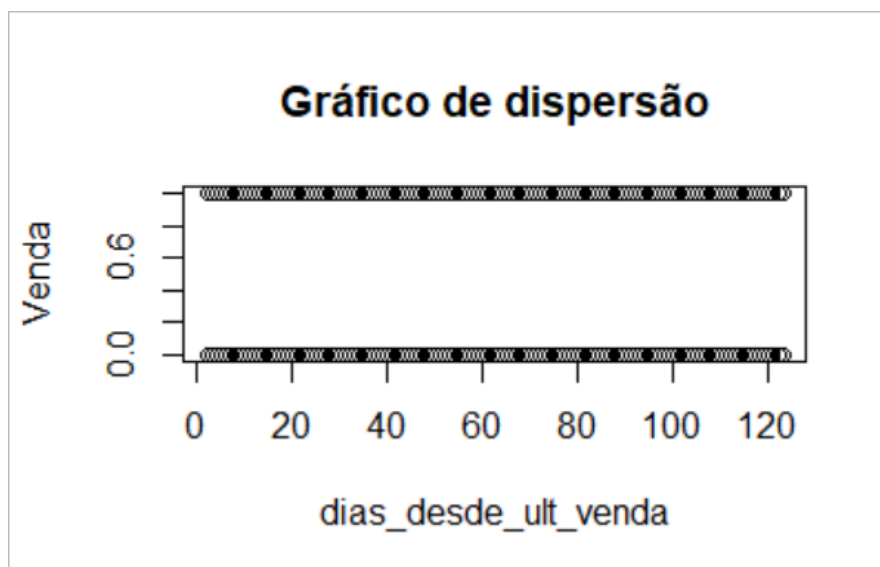


Fonte: Elaborado pelo autor.

Assim, observou-se que vendedores que venderam (venda = 0) tendem a ter uma proporção ativação meses mais alta do que vendedores que não venderam (venda = 1) assim, observa-se que essa variável deve ser interessante para o modelo, contradizendo o gráfico de dispersão.

A variável dias desde a última venda não contém valores faltantes. No Gráfico 4, é apresentado o gráfico de dispersão com a variável resposta. Pode se observar que a ocorrência de valores 0 (zero) ou 1 (um) não está relacionado aos dias desde a última venda.

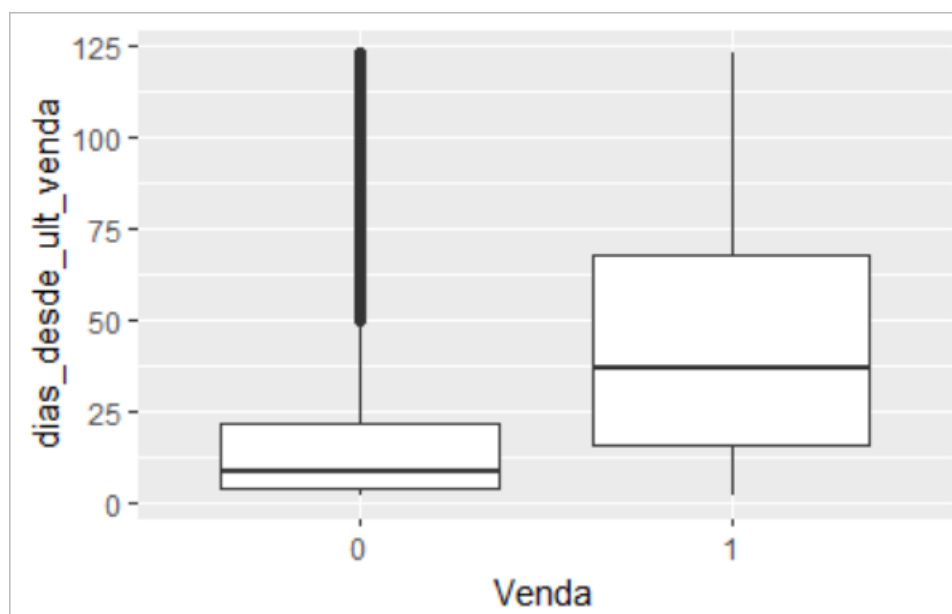
Gráfico 4 – Gráfico de dispersão das variáveis dias desde a última venda e venda



Fonte: Elaborado pelo autor.

Observou-se que a análise visual não gera conclusões claras a partir deste gráfico, então acredita-se que não deve ser uma boa variável para o modelo, mas precisamos observar o boxplot para as conclusões serem mais precisas. No Gráfico 5, é apresentado o Boxplot dos dias desde a última venda entre os vendedores com venda e sem venda.

Gráfico 5 – Boxplot dias desde a última venda por venda



Fonte: Elaborado pelo autor.

Assim, observou-se que vendedores que venderam (venda = 0) tendem a demorar menos dias para vender do que vendedores que não venderam (venda = 1) portanto, essa variável parece interessante para o modelo.

A variável quantidade de produtos no último mês mostra a quantidade de produtos vendidos no último mês antes da coleta atual, ou seja, se o vendedor não efetuou venda há 45 dias, essa variável vai estar com o valor 15, pois os 30 dias passados são para a coleta da variável resposta. Essa variável não contém valores faltantes, notou-se que a variável se comporta de forma que poucos vendedores vendem muitos produtos e muitos vendedores vendem poucos produtos.

No Gráfico 6, é apresentado o gráfico de dispersão com a variável resposta. Pode se observar que a ocorrência de valores 0 (zero) ou 1 (um) reforça a afirmação anterior sobre o comportamento distinto quando há venda e quando não há venda.

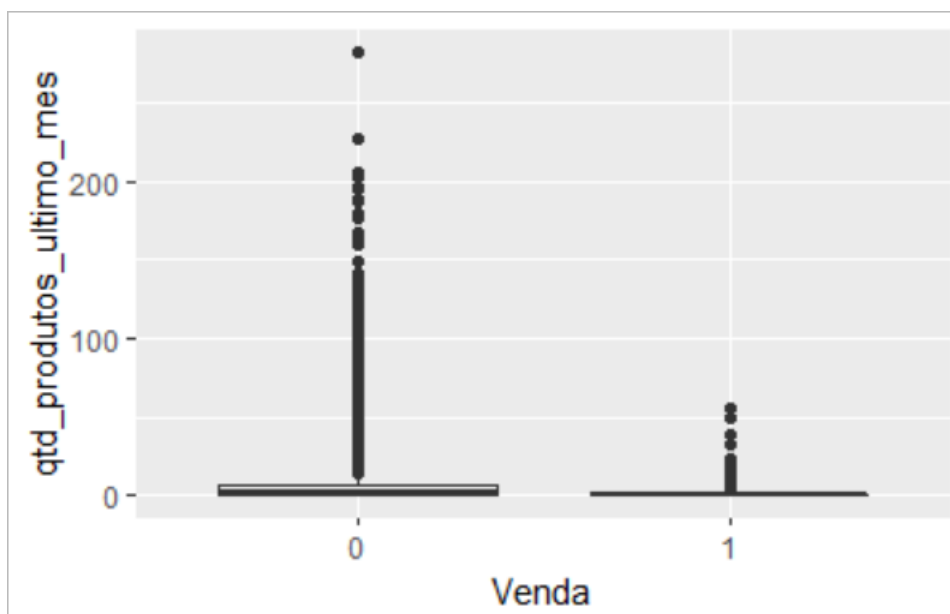
Gráfico 6 – Gráfico de dispersão das variáveis quantidade de produtos no último mês e venda



Fonte: Elaborado pelo autor.

Observou-se que aparentemente a não venda (venda = 1) contempla vendedores que possuem poucos produtos vendidos no último mês, ao observar o boxplot as conclusões são reforçadas. No Gráfico 7, é apresentado o Boxplot da quantidade de produtos no último mês entre os vendedores com venda e sem venda.

Gráfico 7 – Boxplot quantidade de produtos no último mês por venda

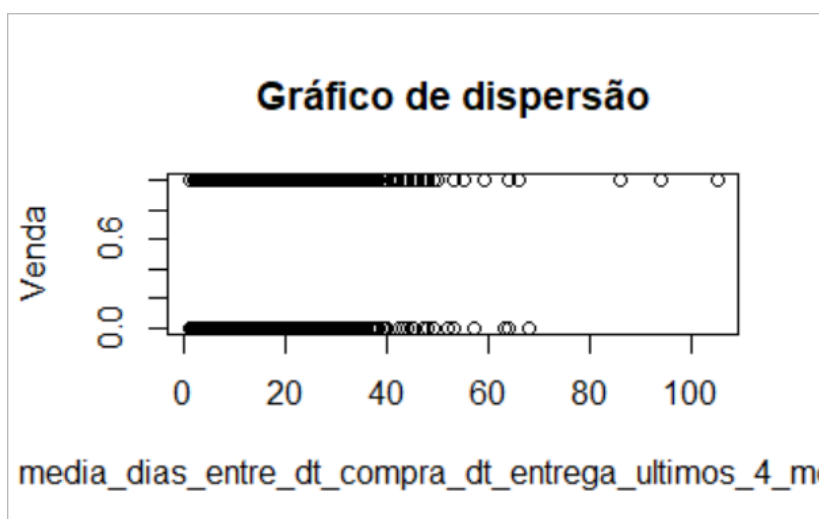


Fonte: Elaborado pelo autor.

O Boxplot mostra que o comportamento é parecido, mas também dá indícios de que vendedores que não efetuaram vendas (venda = 1) tendem a ter uma quantidade de produtos menor no último mês.

A variável média de dias entre a data de compra e a data de entrega nos últimos quatro meses continha 1197 valores faltantes que foram estimados através da média. No Gráfico 8, é apresentado o gráfico de dispersão com a variável resposta. Pode se observar que a ocorrência de valores 0 (zero) ou 1 (um) não está relacionado a média de dias entre a data de compra e data de entrega nos últimos 4 meses.

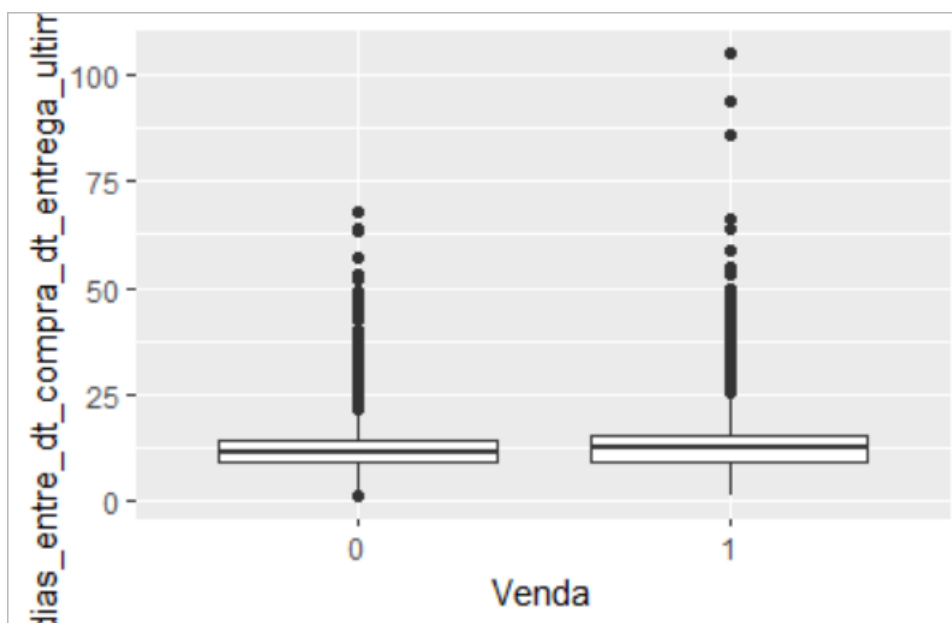
Gráfico 8 – Gráfico de dispersão das variáveis média dias entre dt compra dt entrega últimos 4 meses e venda



Fonte: Elaborado pelo autor.

Observou-se que a análise visual não gera conclusões claras a partir deste gráfico, o que gera preocupações se essa variável será boa para o modelo, mas ao observar o boxplot espera-se que as conclusões sejam mais intuitivas. No Gráfico 9, é apresentado o Boxplot da média de dias entre a data de compra e a data de entrega, nos últimos 4 meses, entre os vendedores com venda e sem venda.

Gráfico 9 – Boxplot media dias entre dt compra dt entrega últimos 4 meses por venda

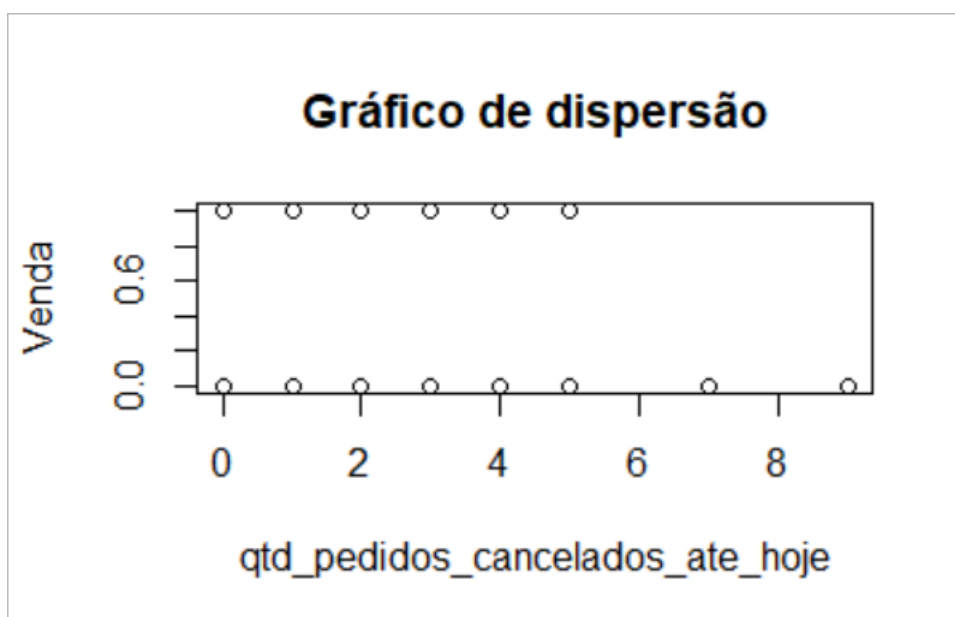


Fonte: Elaborado pelo autor.

O Boxplot mostra que o comportamento é parecido, contém muitos outliers nos dois casos de venda, com isso, essa variável não aparenta ser tão importante quanto as anteriores para prever a venda.

A variável quantidade de pedidos cancelados até hoje não contém valores faltantes. No Gráfico 10, é apresentado o gráfico de dispersão com a variável resposta. Pode se observar que a ocorrência de valores 0 (zero) ou 1 (um) não está relacionado a quantidade de pedidos cancelados até hoje.

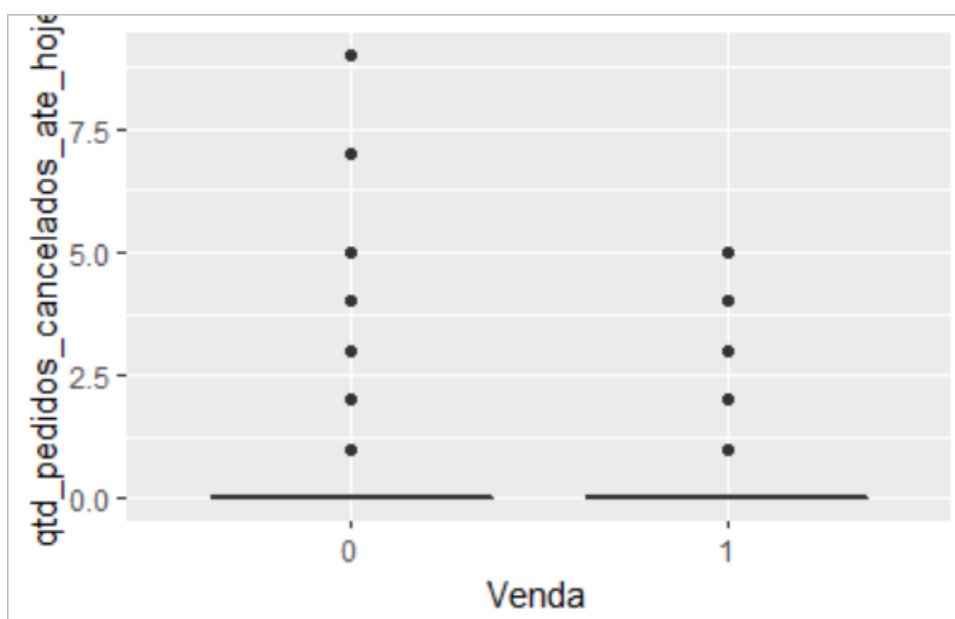
Gráfico 10 – Gráfico de dispersão das variáveis quantidade de pedidos cancelados até hoje e venda



Fonte: Elaborado pelo autor.

Observou-se que a análise visual não gera conclusões claras a partir deste gráfico, o que gera indícios de que essa variável pode não ser interessante para o modelo, mas ao observar o boxplot espera-se que as conclusões fiquem mais intuitivas. No Gráfico 11, é apresentado o Boxplot da quantidade de pedidos cancelados até hoje entre os vendedores com venda e sem venda.

Gráfico 11 – Boxplot quantidade de pedidos cancelados até hoje por venda

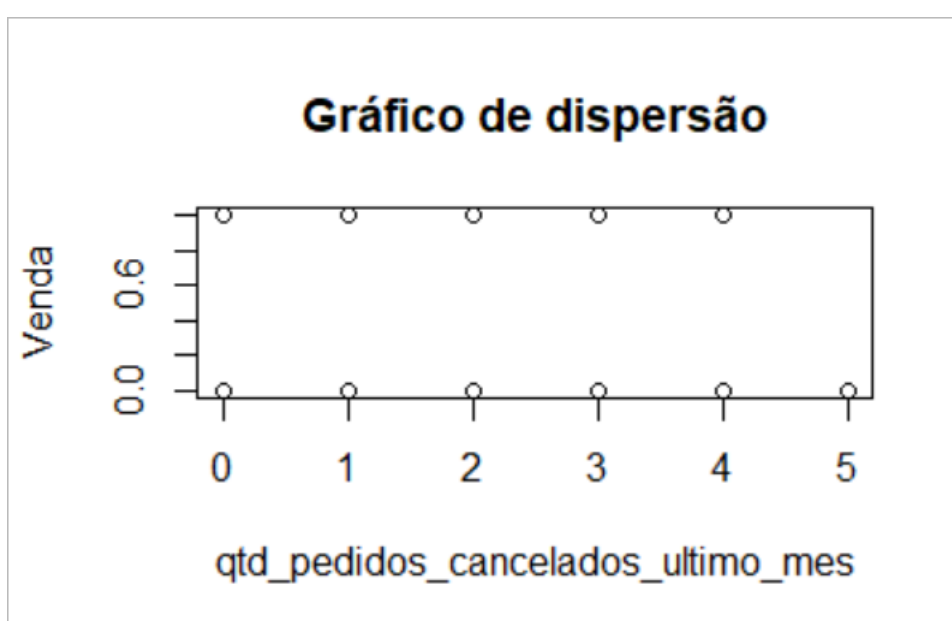


Fonte: Elaborado pelo autor.

O Boxplot mostra que o comportamento é parecido, contém outliers nos dois casos de venda, mas a concentração da quantidade de pedidos cancelados é predominantemente zero para os dois casos.

A variável quantidade de pedidos cancelados último mês não contém valores faltantes. No Gráfico 12, é apresentado o gráfico de dispersão com a variável resposta. Pode se observar que a ocorrência de valores 0 (zero) ou 1 (um) não está relacionado a quantidade de pedidos cancelados no último mês.

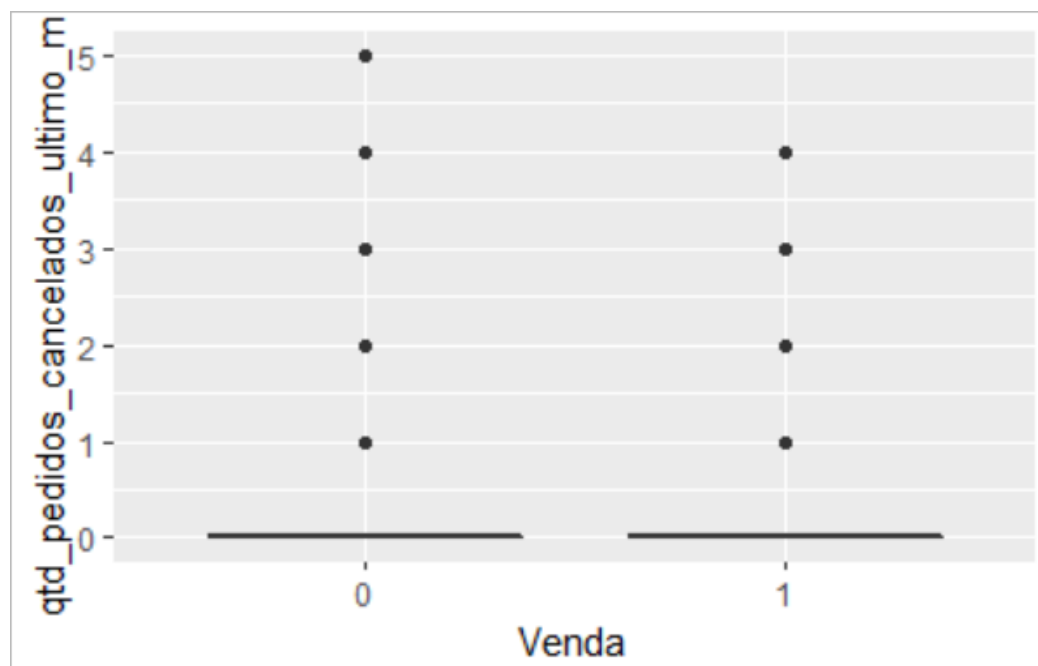
Gráfico 12 – Gráfico de dispersão das variáveis quantidade de pedidos cancelados último mês e venda



Fonte: Elaborado pelo autor.

Observou-se que a análise visual não gera conclusões claras a partir deste gráfico, o que gera indícios de que a variável não deve ser interessante para o modelo, mas ao observar o boxplot espera-se que as conclusões fiquem mais intuitivas. No Gráfico 13, é apresentado o Boxplot da quantidade de pedidos cancelados no último mês entre os vendedores com venda e sem venda.

Gráfico 13 – Boxplot quantidade de pedidos cancelados último mês por venda

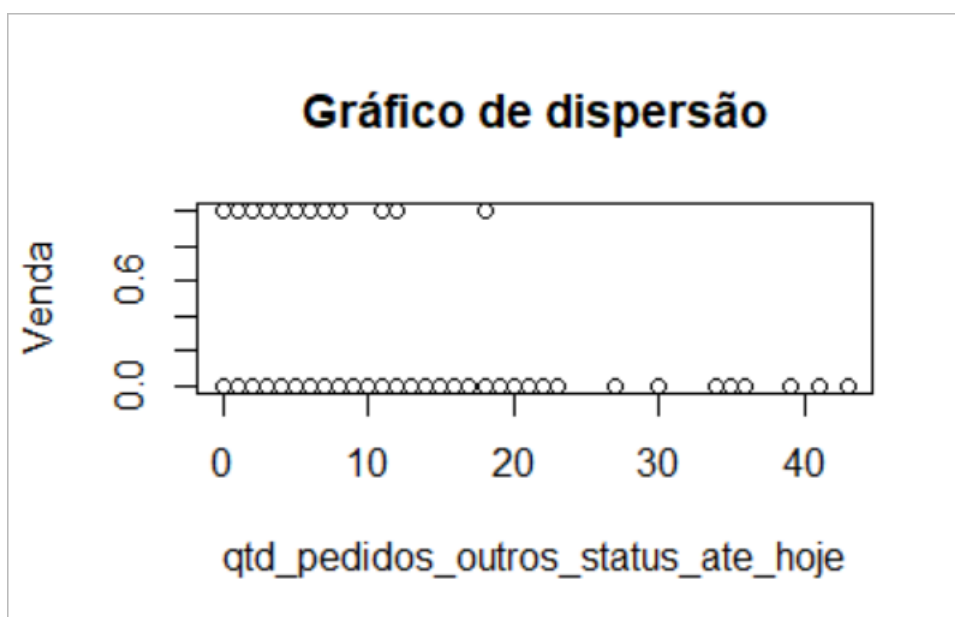


Fonte: Elaborado pelo autor.

O Boxplot mostra que o comportamento é parecido, contém outliers nos dois casos de venda, mas a concentração da quantidade de pedidos cancelados é predominantemente zero para os dois casos, assim reforçando a hipótese de que essa variável não deve ser interessante para o modelo.

A variável quantidade de pedidos outros status até hoje é uma variável que mostra o número de pedidos extraviados, não enviados, atrasados, etc. Essa variável não contém valores faltantes. No Gráfico 14, é apresentado o gráfico de dispersão com a variável resposta. Pode se observar que a ocorrência de valores 0 (zero) ou 1 (um) mostra um comportamento diferente quando há e quando não há venda.

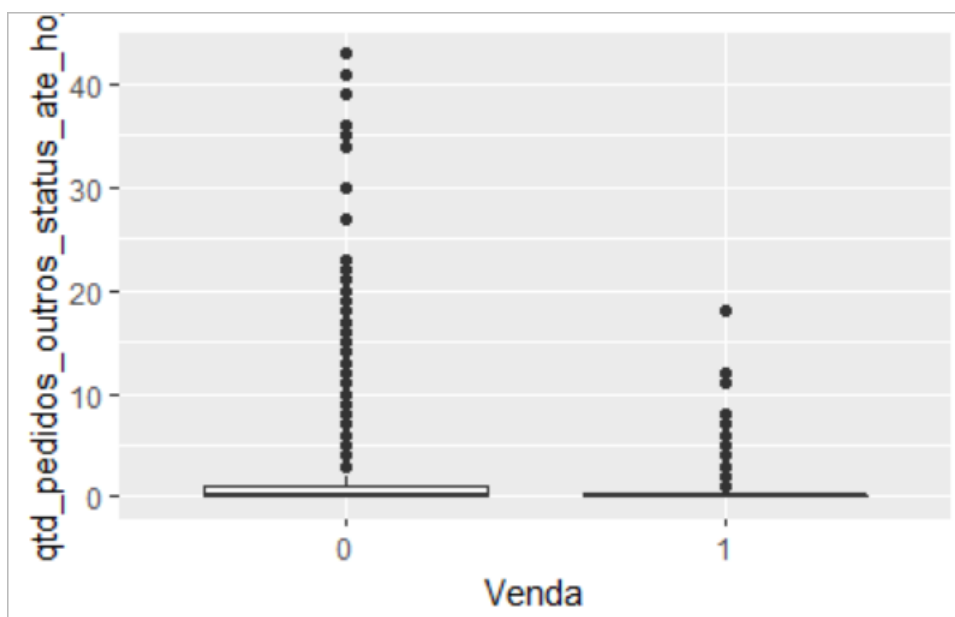
Gráfico 14 –Gráfico de dispersão das variáveis quantidade de pedidos outros status até hoje e venda



Fonte: Elaborado pelo autor.

Observou-se na análise visual que aparentemente os vendedores que venderam (venda = 0) têm mais pedidos em outros status, ou seja, essa variável aparenta ser interessante para o modelo, ao observar o boxplot espera-se confirmar essa suposição. No Gráfico 15, é apresentado o Boxplot da quantidade de pedidos em outros status até hoje, entre os vendedores com venda e sem venda.

Gráfico 15 – Boxplot quantidade de pedidos outros status até hoje por venda

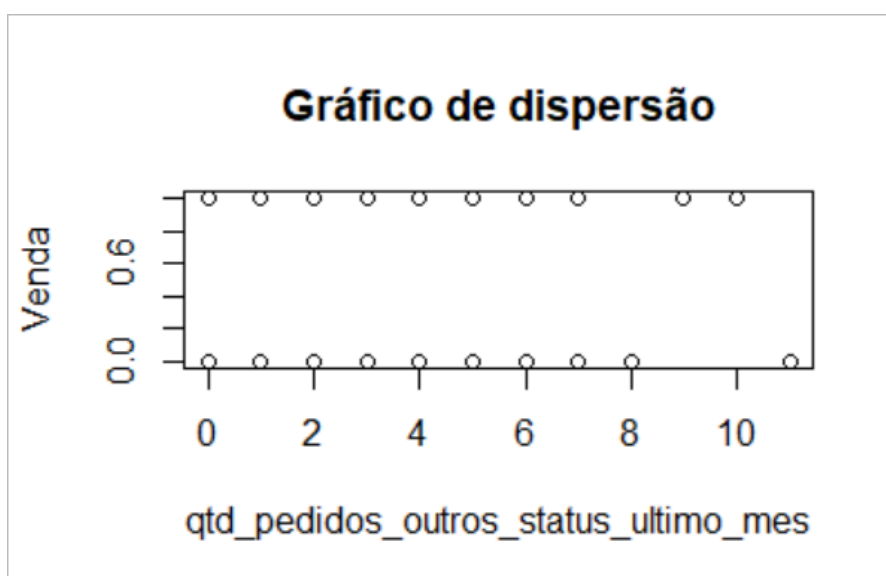


Fonte: Elaborado pelo autor.

O Boxplot reforça a conclusão obtida no gráfico de dispersão, contém outliers nos dois casos de venda, mas a concentração da quantidade de pedidos outros status é predominantemente zero para os dois casos.

A variável quantidade de pedidos outros status último mês é uma variável que mostra o número de pedidos extraviados, não enviados, atrasados, etc. Essa variável não contém valores faltantes. No Gráfico 16, é apresentado o gráfico de dispersão com a variável resposta. Pode se observar que a ocorrência de valores 0 (zero) ou 1 (um) não está relacionado a quantidade de pedidos outros status no último mês.

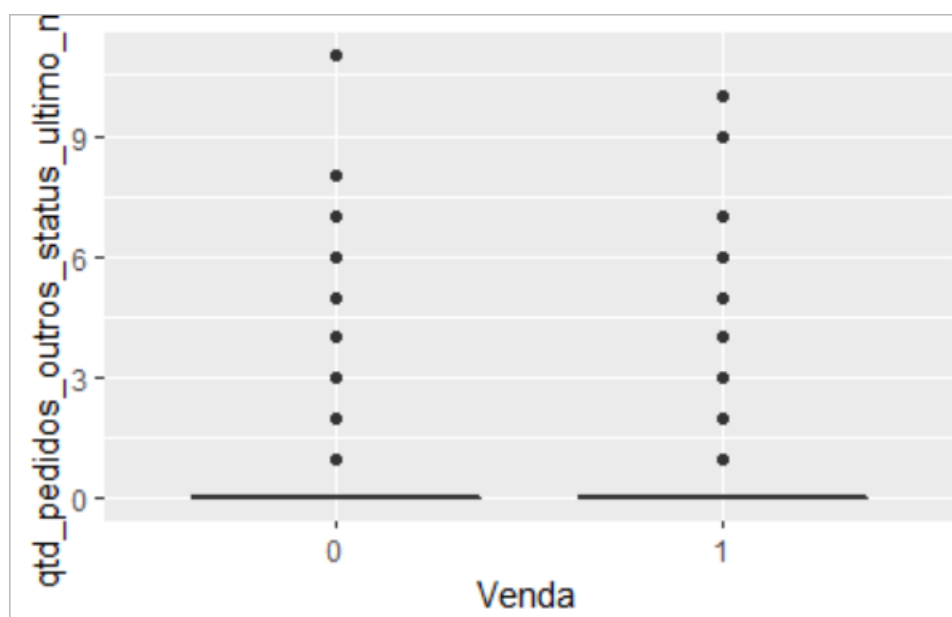
Gráfico 16 –Gráfico de dispersão das variáveis quantidade de pedidos outros status último mês e venda



Fonte: Elaborado pelo autor.

Observou-se que a análise visual não gera conclusões claras a partir deste gráfico, ou seja, espera-se que essa variável não seja interessante para o modelo, mas ao observar o boxplot espera-se que as conclusões fiquem mais intuitivas. No Gráfico 17, é apresentado o Boxplot da quantidade de pedidos outros status no último mês entre os vendedores com venda e sem venda.

Gráfico 17 – Boxplot quantidade de pedidos outros status último mês por venda

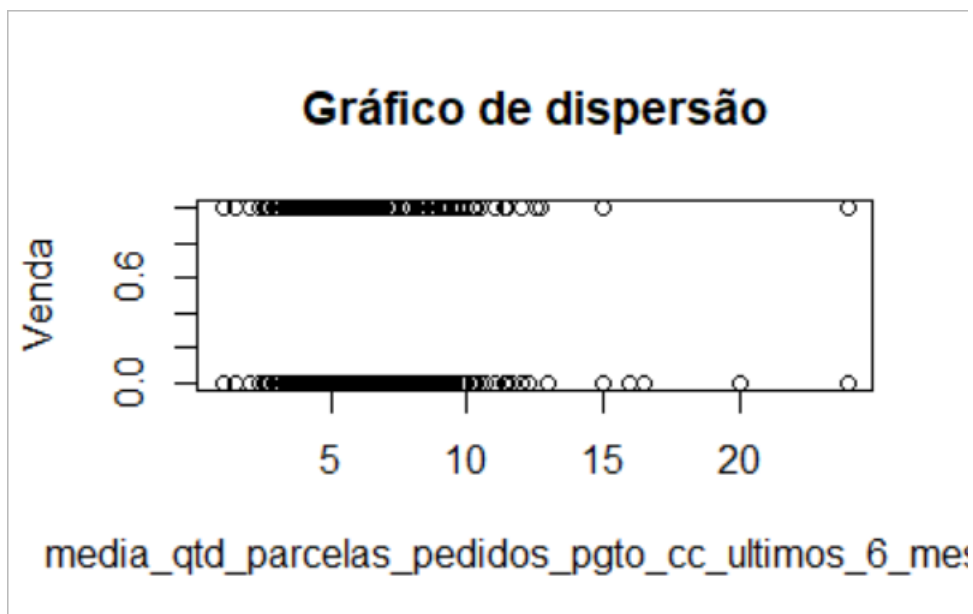


Fonte: Elaborado pelo autor.

O Boxplot mostra que o comportamento é parecido, contém outliers nos dois casos de venda, mas a concentração da quantidade de pedidos outros status é predominantemente zero para os dois casos, o que reforça a hipótese anterior.

A variável média da quantidade de parcelas de pedidos com pagamento no cartão de crédito nos últimos seis meses não contém valores faltantes. No Gráfico 18, é apresentado o gráfico de dispersão com a variável resposta. Pode se observar que a ocorrência de valores 0 (zero) ou 1 (um) não está relacionado a média da quantidade de parcelas em pedidos pagos no cartão de crédito, nos últimos 6 meses.

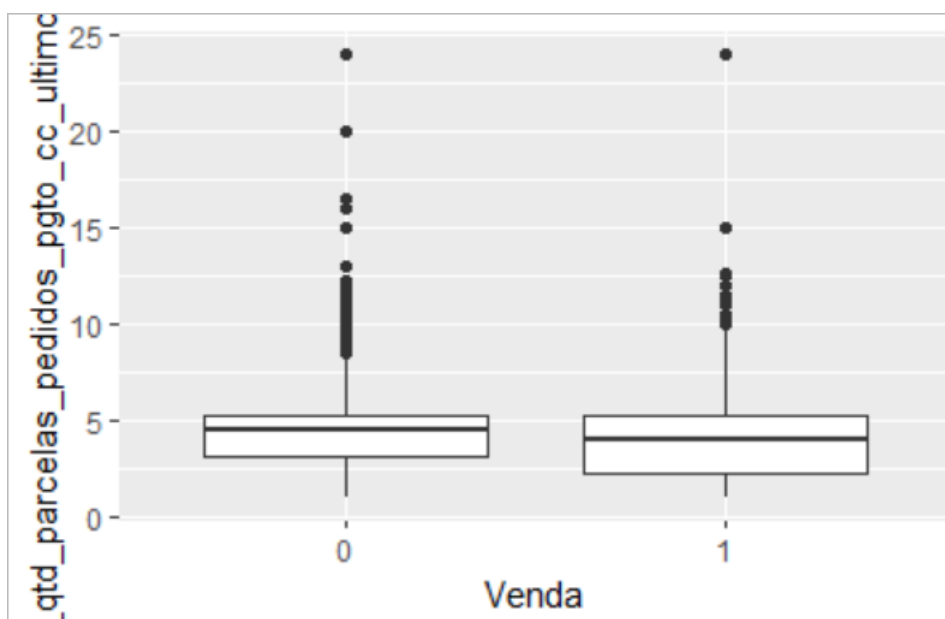
Gráfico 18 –Gráfico de dispersão das variáveis media qtd parcelas pedidos pgto cc últimos 6 meses e venda



Fonte: Elaborado pelo autor.

Observou-se que a análise visual não gera conclusões claras a partir deste gráfico, ou seja, espera-se que essa variável não seja interessante para o modelo, mas ao observar o boxplot espera-se que as conclusões fiquem mais intuitivas. No Gráfico 19, é apresentado o Boxplot da média da quantidade de parcelas para pedidos no cartão de crédito, nos últimos 6 meses, entre os vendedores com venda e sem venda.

Gráfico 19 – Boxplot media qtd parcelas pedidos pgto cc últimos 6 meses por venda

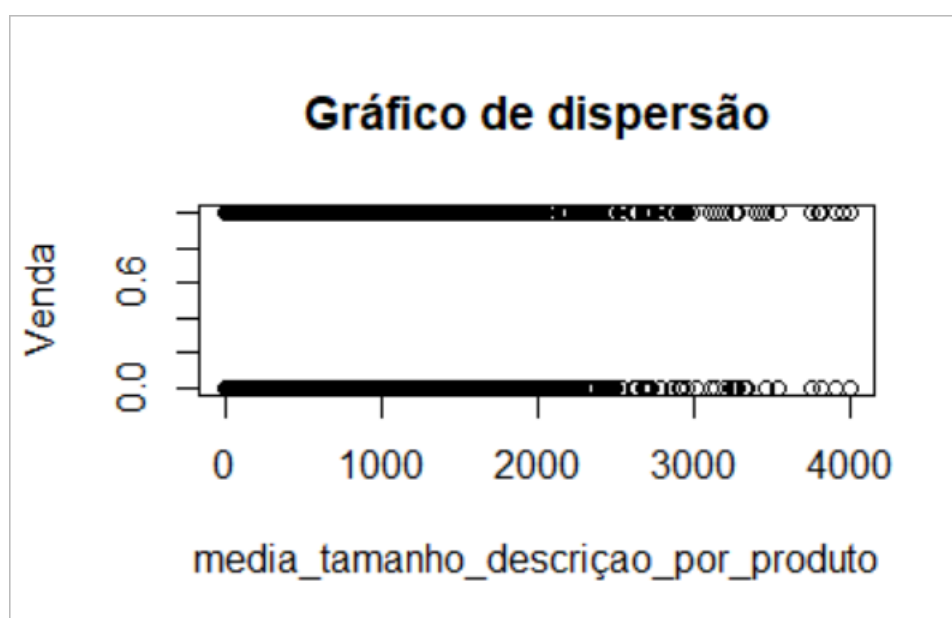


Fonte: Elaborado pelo autor.

O Boxplot mostra que o comportamento é parecido, contém outliers nos dois casos de venda, mas a concentração da média da quantidade de parcelas de pedidos com pagamento no cartão de crédito nos últimos seis meses é próxima de 4, o que reforça a hipótese anterior.

A variável média do tamanho da descrição por produto continha 254 valores faltantes que foram preenchidos com o valor zero. No Gráfico 20, é apresentado o gráfico de dispersão com a variável resposta. Pode se observar que a ocorrência de valores 0 (zero) ou 1 (um) não está relacionado a média do tamanho da descrição por produto.

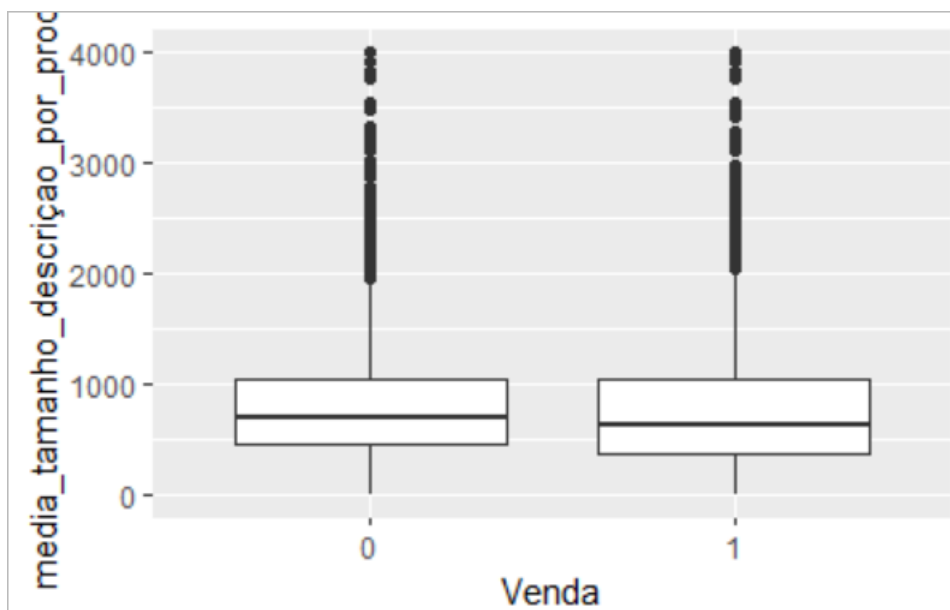
Gráfico 20 –Gráfico de dispersão das variáveis média do tamanho da descrição por produto e venda



Fonte: Elaborado pelo autor.

Observou-se que a análise visual não gera conclusões claras a partir deste gráfico, assim, espera-se que essa variável não seja interessante para o modelo, mas ao observar o boxplot espera-se que as conclusões fiquem mais intuitivas. No Gráfico 21, é apresentado o Boxplot da média do tamanho da descrição entre os vendedores com venda e sem venda.

Gráfico 21 – Boxplot média do tamanho da descrição por produto por venda

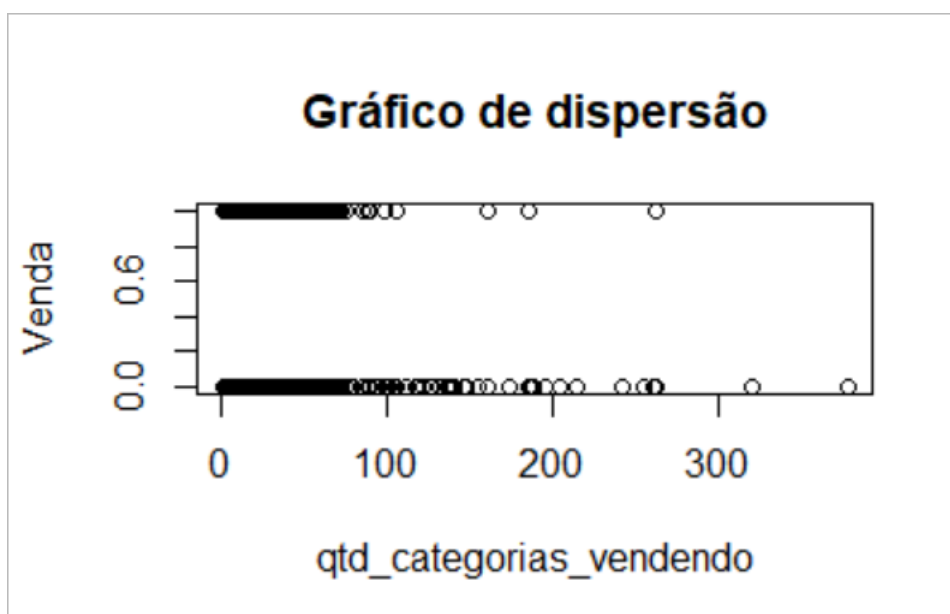


Fonte: Elaborado pelo autor.

O Boxplot mostra que o comportamento é parecido e contém outliers nos dois casos de venda.

A variável quantidade de categorias vendendo é a quantidades de categorias que o vendedor atua, essa variável não contém valores faltantes. No Gráfico 22, é apresentado o gráfico de dispersão com a variável resposta. Pode se observar que a ocorrência de valores 0 (zero) ou 1 (um) mostra um comportamento diferente quando há e quando não há venda.

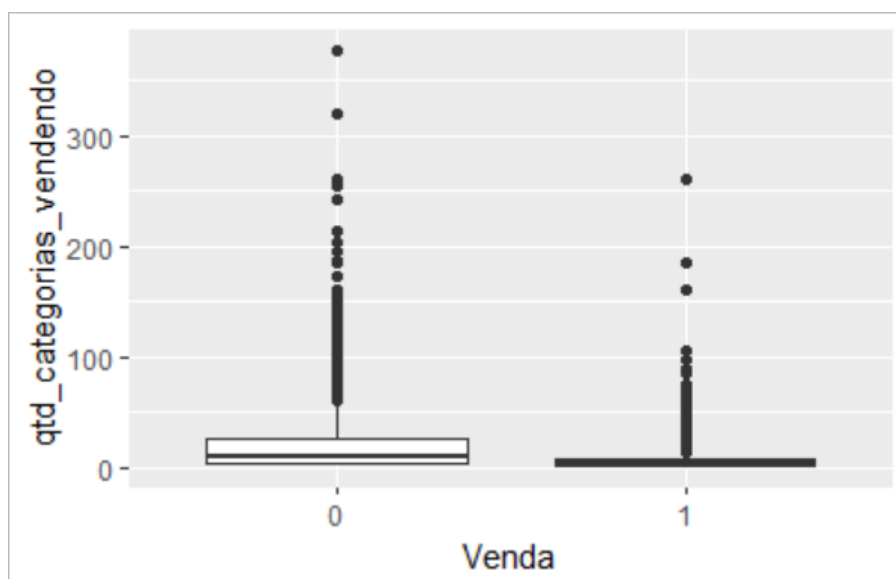
Gráfico 22 –Gráfico de dispersão das quantidades de categorias vendendo e venda



Fonte: Elaborado pelo autor.

Observou-se na análise visual que aparentemente os vendedores que efetuaram venda (venda = 0) tendem a ter mais categorias vendendo, ou seja, essa variável parece ser interessante para o modelo, mas para comprovar isso usou-se o boxplot e espera-se que as conclusões sejam reforçadas. No Gráfico 23, é apresentado o Boxplot da quantidade de categorias vendendo entre os vendedores com venda e sem venda.

Gráfico 23 – Boxplot quantidade de categorias vendendo por venda



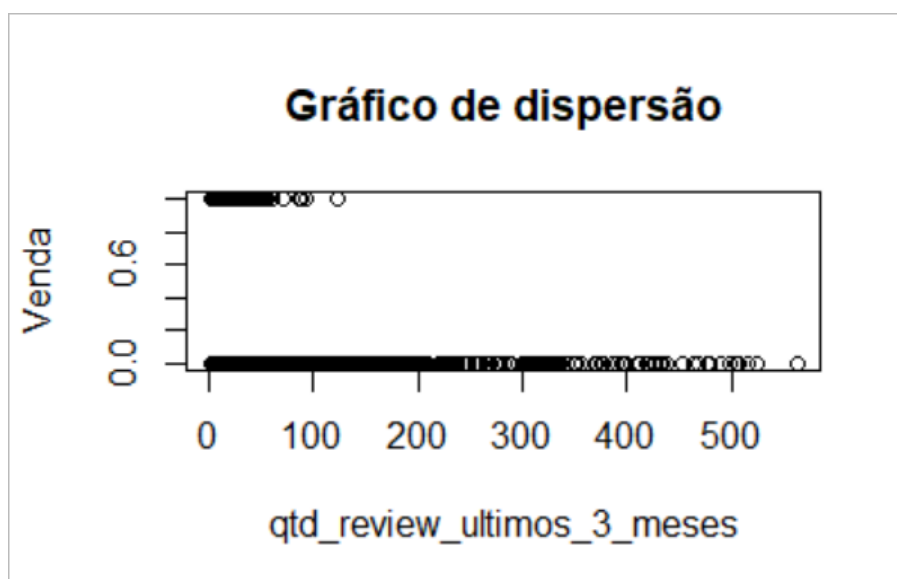
Fonte: Elaborado pelo autor.

Assim, observou-se que vendedores que venderam (venda = 0) tendem a ter uma quantidade maior de categorias vendendo do que vendedores que não venderam (venda = 1), notou-se também que há muitos outliers em ambos os casos.

A variável quantidade de review¹³ nos últimos três meses continha 1070 valores faltantes que foram preenchidos a mediana. No Gráfico 24, é apresentado o gráfico de dispersão com a variável resposta. Pode se observar que a ocorrência de valores 0 (zero) ou 1 (um) mostra um comportamento diferente quando há e quando não há venda.

¹³ Review é o ato de avaliar uma compra feita na internet.

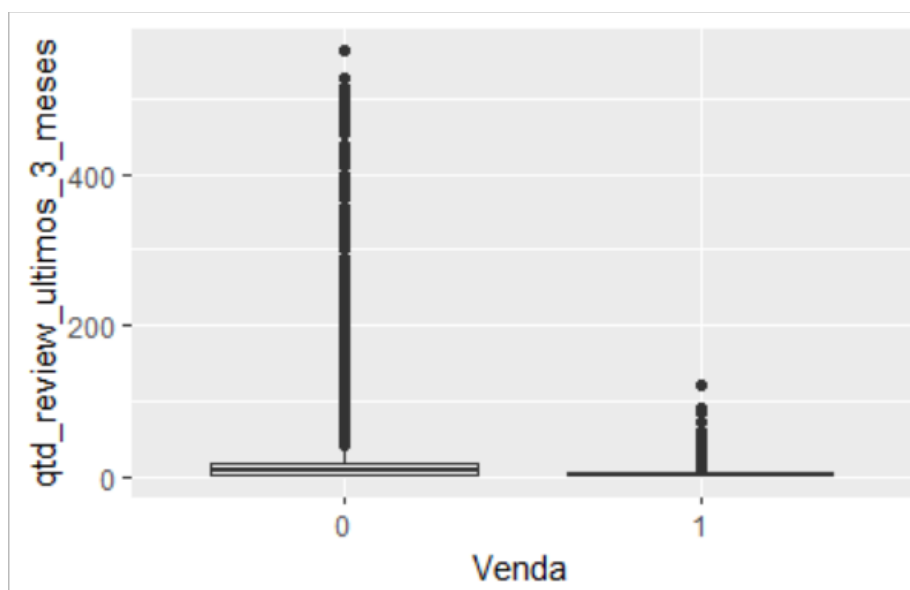
Gráfico 24 –Gráfico de dispersão das variáveis quantidade de review nos últimos 3 meses e venda



Fonte: Elaborado pelo autor.

Observou-se através da análise visual que aparentemente o vendedor que não vendeu (venda = 1) tende a ter menos reviews nos últimos 3 meses do que vendedores que venderam nos últimos 3 meses, mas ao observar o boxplot espera-se confirmar essas afirmações. No Gráfico 25, é apresentado o Boxplot da quantidade de review nos últimos três meses, entre os vendedores com venda e sem venda.

Gráfico 25 – Boxplot quantidade de review nos últimos três meses por venda

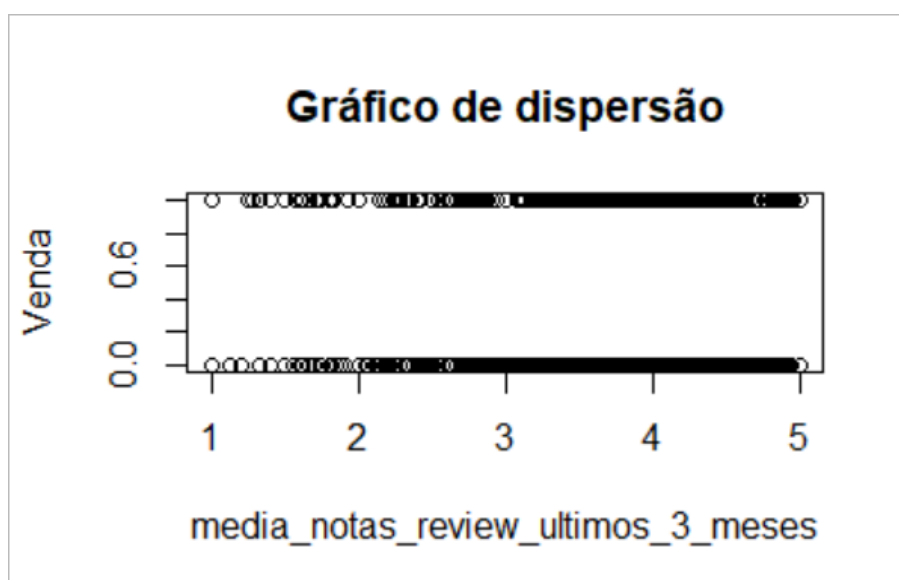


Fonte: Elaborado pelo autor.

O Boxplot confirma a afirmação, mas mostra que possuem vários outliers e que a variável está concentrada próxima de zero, assim, acredita-se que essa variável seja interessante para o modelo.

A variável média das notas de review nos últimos três meses continha 1070 valores faltantes que foram preenchidos a mediana. No Gráfico 26, é apresentado o gráfico de dispersão com a variável resposta. Pode se observar que a ocorrência de valores 0 (zero) ou 1 (um) não está relacionado a média das notas de review nos últimos três meses.

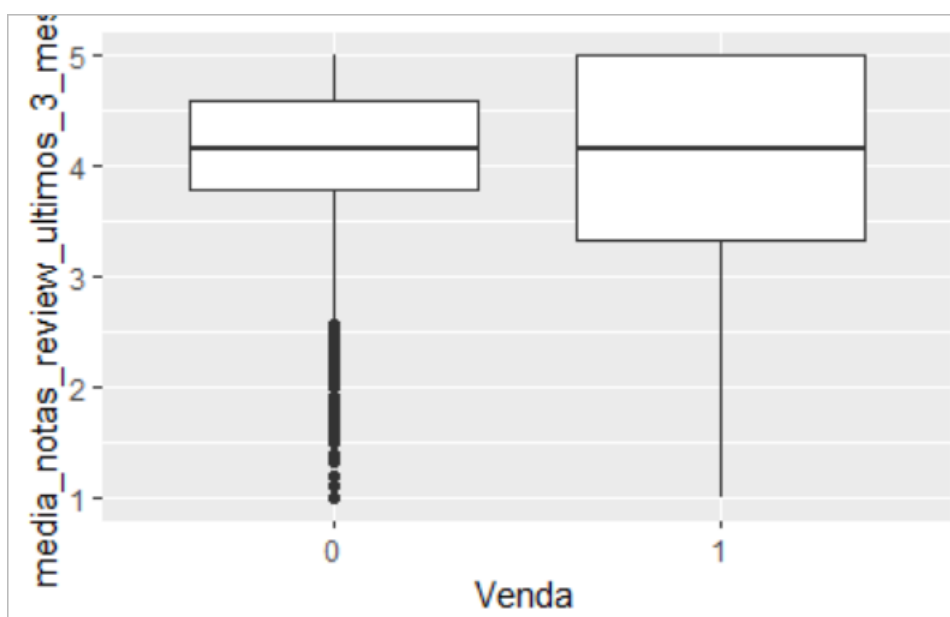
Gráfico 26 –Gráfico de dispersão das variáveis média das notas de review nos últimos três meses e venda



Fonte: Elaborado pelo autor.

Observou-se que a análise visual não gera conclusões claras a partir deste gráfico, ou seja, essa variável não deve ser interessante para o modelo, mas ao observar o boxplot espera-se que as conclusões fiquem mais intuitivas. No Gráfico 27, é apresentado o Boxplot da média das notas de review nos últimos três meses, entre os vendedores com venda e sem venda.

Gráfico 27 – Boxplot média das notas de review nos últimos três meses por venda



Fonte: Elaborado pelo autor.

O Boxplot mostra levemente que as notas médias mais baixas estão levemente mais próximas a vendedores que não efetuaram vendas (venda = 1), assim contrapondo a hipótese de que essa variável não seria interessante.

7.5 Variáveis Dummy

Foi decidido criar variáveis dummy¹⁴ para as variáveis escolhidas após a etapa da análise descritiva, sabe-se também que não poderemos usar todas as variáveis dummy para o modelo, então será escolhida uma das categorias para ficar como referência.

Para uma variável contínua a etapa inicial para a criação de variáveis dummy é a categorização de variáveis, portanto, com ajuda dos conhecimentos obtidos na descritiva foram feitas as criações das classes nas variáveis como mostra a tabela a seguir:

Tabela 1 - Quebras das variáveis

Variável	Quebras	Nomenclaturas
Ativação em meses	0; 0,5; 0,9; 1	Baixa, Média, Alta
Dias desde a última venda	0; 30; 90; ∞	Ate_30, De_30_a_90, Mais_90
Quantidade de produtos último mês	0; 5; 30; ∞	Ate_5, De_5_a_30, Mais_30

¹⁴ Variáveis dummy são n variáveis binárias (0 ou 1) onde n será o número de classes da variável original

Média dias entre data de compra e data de entrega nos últimos 4 meses	0; 9; 15; ∞	Ate_9, De_9_a_15, Mais_15
Quantidade de pedidos cancelados até hoje	0; 0; 1; ∞	Nenhum, Um, Mais_de_1
Quantidade de pedidos cancelados último mês	0; 0; 1; ∞	Nenhum, Um, Mais_de_1
Quantidade de pedidos outros status até hoje	0; 0; 5; ∞	Nenhum, De_1_a_5, Mais_5
Quantidade de pedidos outros status último mês	0; 0; 1; ∞	Nenhum, Um, Mais_de_1
Média da quantidade de parcelas em pedidos pagos no cartão de crédito nos últimos 6 meses	0; 3; 9; ∞	Ate_3, De_3_a_9, Mais_9
Média tamanho da descrição por produto	0; 400; 2000; ∞	Ate_400, De_400_a_2000, Mais_2000
Quantidade de categorias vendendo	0; 5; 10; 60; ∞	Ate_5, De_5_a_10, De_10_a_60, Mais_60
Quantidade de reviews nos últimos 3 meses	0; 1; 6; 20; ∞	Ate_1, De_1_a_6, De_6_a_20, Mais_20

Fonte: Elaborado pelo autor.

Após a separação nas classes foram criadas $n-1$ variáveis binárias para cada variável que foi dividida em classes, onde o n é o número de classes em que foi feita a divisão, ou seja, para a variável ativação em meses foram criadas outras duas variáveis ativação em meses média e ativação em meses alta, a ativação e meses baixa está representada como 0 na ativação média e 0 na ativação alta, desta forma, todas as primeiras classes foram retiradas como referência para o momento da interpretação do modelo.

7.6 Qualidade do ajuste

Antes de começar com as manipulações de dados, os dados foram divididos em treino e teste, a parte de treino ficou com 70% da base de dados (o que representa 13627 linhas) e a base de teste ficou com 30% (5841 linhas). Essa repartição manteve a taxa da variável resposta que é de 0.375.

Como descobriu-se no contato com a área de negócios que o foco do modelo é ordenação, portanto, a principal métrica que foi analisada é o lift, mas para fins didáticos e de aprendizado foi feita a matriz de confusão, bem como, a curva ROC.

Nesta etapa, observa-se o desempenho do modelo de Regressão Logística descrito na seção 7.7, aplicado na base de treino (a base em que ele foi criado).

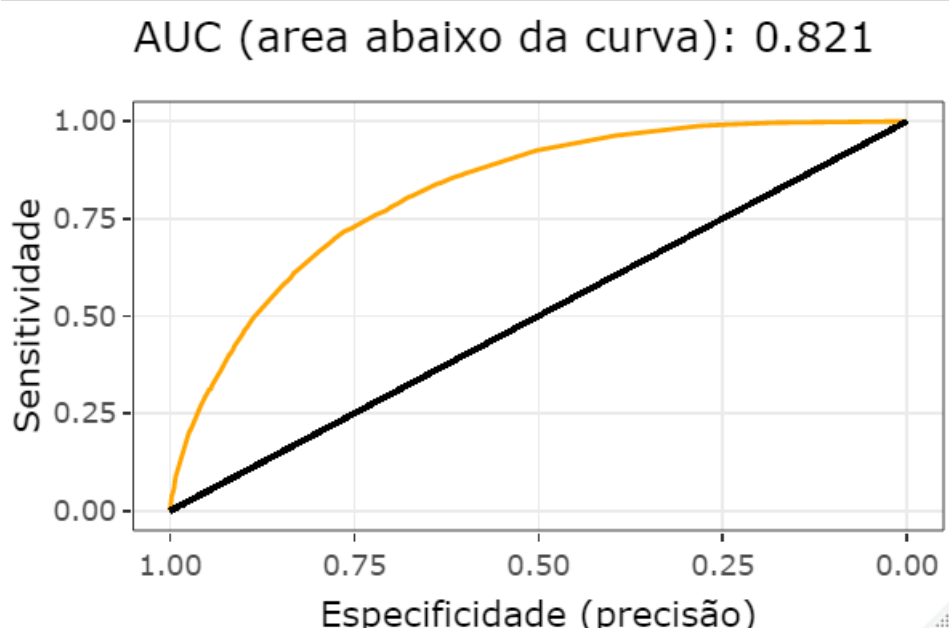
Figura 7 – Matriz de confusão treino

		Previsto (modelo)	
		0	1
Real	0	7617	899
	1	2677	2434

Fonte: Elaborado pelo autor.

A partir da matriz obteve-se as métricas de acurácia, precisão e recall. Acurácia: 0,7375, então a taxa de acerto geral é de 0,7375 ou 73,75%. Precisão: 0,73, logo a porcentagem de eventos de interesse que o modelo acertou é de 73%. Recall: 0,4762, assim a porcentagem de eventos de interesse capturada pelo modelo é de 47,62%. Vale lembrar que a relação entre a precisão e o recall é inversamente proporcional, ou seja, se o ponto de corte (que atualmente é 0.6) for mudado e o recall subir, a precisão vai diminuir, caso o recall diminua, então a precisão aumenta.

Gráfico 28 – Curva ROC treino



Fonte: Elaborado pelo autor.

A curva ROC mostra-se interessante, aparenta ser um modelo bom e a área abaixo da curva (conhecido como AUC) é de 0,821, o que parece ser bom, já que quanto mais próximo de 1 melhor.

Finalmente veremos a métrica lift. Para ela foram criados dois cenários, o primeiro cenário em que a área de marketing da olist tem orçamento para abordar 10% da base de vendedores, já o segundo supõe que a área tenha orçamento para abordar 20% dos vendedores.

Para o primeiro cenário tem-se um lift de 2,1654, ou seja, pode-se dizer para a área de marketing que o modelo é 116,54% (ou 1,1654) vezes melhor do que escolher aleatoriamente os indivíduos. Outra forma de explicar essa informação é dizer que utilizando o modelo atual a área de marketing terá 116,54% mais retorno na campanha.

Já para o segundo cenário tem-se um lift de 2,0011, assim, dizer que o modelo é 100,11% (ou 1,0011) vezes melhor do que escolher aleatoriamente os indivíduos. Também pode-se dizer que a área de marketing terá 100,11% mais retorno na campanha ao utilizar o modelo proposto.

Portanto, o ajuste do modelo aparenta estar bom, assim segue-se para a interpretação do modelo.

7.7 Modelo de Regressão Logística

O modelo foi criado inicialmente com todas as variáveis dummy geradas, mas após isso foi utilizado um método para seleção de variáveis, nesse caso o stepwise que acabou retirando algumas variáveis dummy que eram não significativas no teste Z-Wald.

Tabela 2 – Modelo completo

Variável	Valor estimado	P-valor do teste Z-Wald
Intercepto	0,88774	2,25e-15
Proporção ativação meses Media	-0,08097	0,244989
proporcao ativacao meses Alta	-0,3113	8,73e-06
dias desde ult venda De 30 a 90	0,66210	2e-16
dias desde ult venda Mais 90	1,66613	2e-16
qtd produtos ultimo mes De 5 a 30	-1,42602	2e-16
qtd produtos ultimo mes Mais 30	-3,20209	2,51e-05
media dias entre dt compra dt entrega ultimos 4 meses De 9 a 15	0,12378	0,015227
media dias entre dt compra dt entrega ultimos 4 meses Mais 15	0,23140	0,000374
qtd pedidos cancelados ate hoje Um	0,43481	2,54e-05
qtd pedidos cancelados ate hoje Mais de 1	0,31031	0,163059

qtd pedidos cancelados ultimo mes Um	0,53291	0,012180
qtd pedidos cancelados ultimo mes Mais de 1	1,28509	0,108635
qtd pedidos outros status ate hoje De 1 a 5	0,31207	4,10e-06
qtd pedidos outros status ate hoje Mais 5	0,47876	0,138763
qtd pedidos outros status ultimo mes Um	0,14434	0,284227
qtd pedidos outros status ultimo mes Mais de 1	1,50689	3,68e-06
qtd pedidos dt prevista menor data entrega ate hoje Um	-0,19230	0,002423
qtd pedidos dt prevista menor data entrega ate hoje Mais de 1	-0,27721	0,000487
media qtd parcelas pedidos pgto cc ultimos 6 meses De 3 a 9	-0,06383	0,167417
media qtd parcelas pedidos pgto cc ultimos 6 meses Mais 9	0,28450	0,010551
media tamanho descricao por produto De 400 a 2000	-0,29514	8,30e-09
media tamanho descricao por produto Mais 2000	-0,21164	0,052656
qtd categorias vendendo De 5 a 10	-0,27314	7,03e-07
qtd categorias vendendo De 10 a 60	-0,81036	2e-16
qtd categorias vendendo Mais 60	-1,87551	2,53e-12
qtd review ultimos 3 meses De 1 a 6	-0,51164	2e-16
qtd review ultimos 3 meses De 6 a 20	-0,98393	2e-16
qtd review ultimos 3 meses Mais 20	-1,64162	2e-16
media notas review ultimos 3 meses De 2 a 3	-0,09001	0,396668
media notas review ultimos 3 meses De 3 a 4	-0,36803	7,36e-05
media notas review ultimos 3 meses Mais 4	-0,49429	2,50e-08

Fonte: Elaborado pelo autor.

Agora o modelo final ficou com as seguintes variáveis:

Tabela 3 – Modelo Final

Variável	Valor estimado	P-Valor do teste Z-Wald
Intercepto	0,79235	2e-16
proporcao ativacao meses Alta	-0,26120	5,86e-07
dias desde ult venda De 30 a 90	0,67390	2e-16
dias desde ult venda Mais 90	1,70473	2e-16
qtd produtos ultimo mes De 5 a 30	-1,42146	2e-16
qtd produtos ultimo mes Mais 30	-3,06046	4,84e-05
media dias entre dt compra dt entrega ultimos 4 meses De 9 a 15	0,12115	0,017239
media dias entre dt compra dt entrega ultimos 4 meses Mais 15	0,22556	0,000511
qtd pedidos cancelados ate hoje Um	0,41146	5,97e-05
qtd pedidos cancelados ultimo mes Um	0,58401	0,005553
qtd pedidos cancelados ultimo mes Mais de 1	1,65191	0,033669
qtd pedidos outros status ate hoje De 1 a 5	0,33659	4,35e-08
qtd pedidos outros status ultimo mes Mais de 1	1,52763	1,84e-06
qtd pedidos dt prevista menor data entrega ate hoje Um	-0,20012	0,001512
qtd pedidos dt prevista menor data entrega ate hoje Mais de 1	-0,28728	0,000254
media qtd parcelas pedidos pgto cc ultimos 6 meses Mais 9	0,32889	0,002332
media tamanho descricao por produto De 400 a 2000	-0,29702	6,41e-09
media tamanho descricao por produto Mais 2000	-0,21622	0,047636
qtd categorias vendendo De 5 a 10	-0,28718	1,42e-07
qtd categorias vendendo De 10 a 60	-0,82004	2e-16
qtd categorias vendendo Mais 60	-1,84780	4,55e-12

qtd review ultimos 3 meses De 1 a 6	-0,54233	2e-16
qtd review ultimos 3 meses De 6 a 20	-1,01215	2e-16
qtd review ultimos 3 meses Mais 20	-1,64475	2e-16
media notas review ultimos 3 meses De 3 a 4	-0,32509	1,38e-06
media notas review ultimos 3 meses Mais 4	-0,45550	2,08e-13

Fonte: Elaborado pelo autor.

Sabe-se que a equação do modelo é a equação (3.3) onde os valores de β_i são os valores apresentados acima, mas escolheu-se representar o modelo em uma tabela para facilitar o entendimento individual de cada variável, além disso, ficaria confuso mostrar toda a equação com os respectivos valores.

Os testes Z-Wald que são os testes individuais de cada variável, mostram que todas as variáveis que ficaram no modelo final são significativas para prever a variável venda.

O objetivo a seguir é observar o intervalo de confiança de cada um dos parâmetros do modelo.

Tabela 4 – Intervalo de confiança dos parâmetros

Variável	Intervalo de Confiança 95%
Intercepto	[0,61625738; 0,968452017]
proporcao ativacao meses Alta	[-0,36367856; -0,158724844]
dias desde ult venda De 30 a 90	[0,56599823; 0,781799566]
dias desde ult venda Mais 90	[1,50657029; 1,902885309]
qtd produtos ultimo mes De 5 a 30	[-1,69794447; -1,144974728]
qtd produtos ultimo mes Mais 30	[-4,53667945; -1,58423055]
media dias entre dt compra dt entrega ultimos 4 meses De 9 a 15	[0,02144614; 0,220862064]
media dias entre dt compra dt entrega ultimos 4 meses Mais 15	[0,09833844; 0,352771762]
qtd pedidos cancelados ate hoje Um	[0,21055711; 0,612363693]
qtd pedidos cancelados ultimo mes Um	[0,17123569; 0,996777306]
qtd pedidos cancelados ultimo mes Mais de 1	[0,12755781; 3,176265993]
qtd pedidos outros status ate hoje De 1 a 5	[0,21611823; 0,457066366]
qtd pedidos outros status ultimo mes Mais de 1	[0,89990554; 2,155353863]
qtd pedidos dt prevista menor data entrega ate hoje Um	[-0,32376203; -0,07648317]
qtd pedidos dt prevista menor data entrega ate hoje Mais de 1	[-0,44120298; -0,133357817]

media qtd parcelas pedidos pgto cc ultimos 6 meses Mais 9	[0,11714462; 0,540643384]
media tamanho descricao por produto De 400 a 2000	[-0,39728854; -0,196746864]
media tamanho descricao por produto Mais 2000	[-0,43018606; -0,002249341]
qtd categorias vendendo De 5 a 10	[-0,39414107; -0,180225730]
qtd categorias vendendo De 10 a 60	[-0,93390658; -0,706177823]
qtd categorias vendendo Mais 60	[-2,37124158; -1,324358223]
qtd review ultimos 3 meses De 1 a 6	[-0,64597864; -0,438680764]
qtd review ultimos 3 meses De 6 a 20	[-1,16852654; -0,855770864]
qtd review ultimos 3 meses Mais 20	[-1,99495598; -1,294548218]
media notas review ultimos 3 meses De 3 a 4	[-0,45706402; -0,193112384]
media notas review ultimos 3 meses Mais 4	[-0,57707824; -0,333928163]

Fonte: Elaborado pelo autor.

O Deviance residual obtido no modelo final foi de 13478,91 com 13601 graus de liberdade.

Um exemplo do uso deste modelo de Regressão Logística foi criado para auxiliar o entendimento da técnica. Foram gerados valores para todas as variáveis de três vendedores fictícios, como mostrado na tabela a seguir:

Tabela 5 – Exemplo do desempenho do modelo

Variável	Vendedor 1	Vendedor 2	Vendedor 3
Proporção de ativação em meses	0,9	0,7	0,3
Dias desde a última venda	3	45	90
Quantidade de produtos no último mês	85	22	1
Media dias entre data de compra e data de entrega nos últimos 4 meses	6	10	22
Quantidade de pedidos cancelados até hoje	0	1	3
Quantidade de pedidos cancelados último mês	0	1	2

Quantidade de pedidos outros status até hoje	0	3	15
Quantidade de pedidos outros status último mês	0	1	6
Quantidade de pedidos data de prevista menor que data de entrega até hoje	156	1	0
Media quantidade de parcelas pedidos pagos no cartão últimos 6 meses	2	6	12
Media tamanho da descrição por produto	3516	1350	265
Quantidade de categorias vendendo	72	7	1
Quantidade de review nos últimos 3 meses	24	5	0
Media notas review últimos 3 meses	4,5	3,2	1,8
Valor previsto:	0,0009286446	0,4616419541	0,9980365234

Fonte: Elaborado pelo autor

Foi analisado o valor previsto de cada indivíduo e observou-se que o vendedor 1 tem perfil de um vendedor mais consolidado, fazendo muitas vendas e tendo um ótimo desempenho, enquanto o vendedor 2 é um vendedor mais mediano, que faz algumas vendas mas não mantém uma constância total nas vendas e, finalmente, o vendedor 3 aparenta um perfil de vendedor que não vende a um tempo e teve problemas com os feedbacks dos clientes.

Portanto, pelo exemplo vemos que o modelo está prevendo de forma correta e o foco está realmente nos vendedores que não efetuarão vendas no mês seguinte.

7.8 Interpretação do modelo

A interpretação do modelo de regressão logística é feita por meio das odds ratio, ou então razão das chances. Desta forma foram feitas todas as odds e interpretadas uma a uma da seguinte maneira.

Tabela 6 – Interpretação das odds ratio

Variável	Valor de e^{β_i}	Interpretação
proporcao ativacao meses Alta	0,7701255	Os vendedores que têm a proporção de ativação em meses acima de 0,9 tem 23% menos chance de não vender do que os outros vendedores.
dias desde ult venda De 30 a 90	1,961872	Os vendedores que efetuaram a última venda entre 30 e 90 dias têm 96% mais chance de não efetuar venda do que os vendedores que venderam nos últimos 30 dias.
dias desde ult venda Mais 90	5,499889	Os vendedores que efetuaram a última venda a mais de 90 dias têm 449% mais chance de não efetuar venda do que os vendedores que venderam nos últimos 30 dias.
qtd produtos ultimo mes De 5 a 30	0,2413615	Os vendedores que venderam entre 5 a 30 produtos no último mês têm 75% menos chance de não efetuar venda do que os vendedores que venderam menos que 5 produtos.
qtd produtos ultimo mes Mais 30	0,04686637	Os vendedores que venderam mais de 30 produtos no último mês têm 95% menos chance de não efetuar venda do que os vendedores que venderam menos que 5 produtos.
media dias entre dt compra dt entrega ultimos 4 meses De 9 a 15	1,128799	Os vendedores que venderam produtos que demoraram entre 9 e 15 dias para entrega têm 12% mais chance de não efetuar venda do que vendedores que venderam produtos com data de entrega menor que 9 dias.
media dias entre dt compra dt entrega ultimos 4 meses Mais 15	1,253018	Os vendedores que venderam produtos que demoraram mais de 15 dias para entrega têm 25% mais chance de não efetuar venda do que vendedores que venderam produtos com data de entrega menor que 9 dias.
qtd pedidos cancelados ate hoje Um	1,50902	Os vendedores com um pedido cancelado até hoje têm 50% mais chance de não efetuar venda no próximo mês em relação a vendedores com nenhum ou mais pedidos cancelados.
qtd pedidos cancelados ultimo mes Um	1,793209	Os vendedores com um pedido cancelado no último mês têm 79% mais chance de não efetuar venda no próximo mês em relação a vendedores com nenhum pedido cancelado no último mês.
qtd pedidos cancelados ultimo mes Mais de 1	5,216945	Os vendedores com mais de um pedido cancelado no último mês têm 421% mais chance de não efetuar venda no próximo mês

		em relação a vendedores com nenhum pedido cancelado no último mês.
qtd pedidos outros status ate hoje De 1 a 5	1,400168	Os vendedores que tiveram entre 1 e 5 pedidos com outros status até hoje têm 40% mais chance de não efetuar venda no próximo mês em relação a vendedores com outras quantidades de pedidos em outros status.
qtd pedidos outros status ultimo mes Mais de 1	4,607243	Os vendedores com mais de 1 pedido em outros status no último mês têm 360% mais chance de não efetuar venda com relação a vendedores sem pedidos em outros status.
qtd pedidos dt prevista menor data entrega ate hoje Um	0,8186304	Os vendedores com um pedido entregue em dia têm 18% menos chance de não efetuar venda do que vendedores com nenhum pedido entregue em dia.
qtd pedidos dt prevista menor data entrega ate hoje Mais de 1	0,7503013	Os vendedores com mais de um pedido entregue em dia têm 24% menos chance de não efetuar venda do que vendedores com nenhum pedido entregue em dia.
media qtd parcelas pedidos pgto cc ultimos 6 meses Mais 9	1,389431	Os vendedores com média de quantidades de parcelas dos pedidos pagos no cartão de credito nos últimos 6 meses maior que 9 têm 38% mais chance de não efetuar venda do que vendedores com média de pedidos pagos no cartão menor que 9.
media tamanho descricao por produto De 400 a 2000	0,7430309	Os vendedores com média de caracteres na descrição por produto de 401 a 2000 têm 25% menos chance de não venderem no próximo mês em relação a vendedores com média de descrições inferiores a 400.
media tamanho descricao por produto Mais 2000	0,8055599	Os vendedores com média de caracteres na descrição por produto maior que 2000 têm 19% menos chance de não venderem no próximo mês em relação a vendedores com media de descrições inferiores a 400.
qtd categorias vendendo De 5 a 10	0,7503741	Os vendedores que têm de 5 a 10 categorias vendendo tem 24% menos chance de não efetuar venda do que vendedores com até 5 categorias vendendo.
qtd categorias vendendo De 10 a 60	0,4404131	Os vendedores que têm de 10 a 60 categorias vendendo tem 55% menos chance de não efetuar venda do que vendedores com até 5 categorias vendendo.
qtd categorias vendendo Mais 60	0,1575835	Os vendedores que têm mais de 60 categorias vendendo tem 84% menos chance de não efetuar venda do que vendedores com até 5 categorias vendendo.
qtd review ultimos 3	0,5813922	Os vendedores que tiveram de 1 a 6 reviews nos últimos 3 meses têm 41% menos chance

meses De 1 a 6		de não efetuar venda do que vendedores com até 1 review nos últimos 3 meses.
qtd review ultimos 3 meses De 6 a 20	0,3634372	Os vendedores que tiveram de 6 a 20 reviews nos últimos 3 meses têm 63% menos chance de não efetuar venda do que vendedores com até 1 review nos últimos 3 meses.
qtd review ultimos 3 meses Mais 20	0,1930604	Os vendedores que tiveram mais de 20 reviews nos últimos 3 meses têm 80% menos chance de não efetuar venda do que vendedores com até 1 review nos últimos 3 meses.
media notas review ultimos 3 meses De 3 a 4	0,7224636	Os vendedores com nota média nos últimos 3 meses entre 3 e 4 têm 27% menos chance de não efetuar venda no próximo mês do que vendedores com nota média até 2 nos últimos 3 meses.
media notas review ultimos 3 meses Mais 4	0,6341288	Os vendedores com nota média maior que 4 nos últimos 3 meses têm 36% menos chance de não efetuar venda no próximo mês do que vendedores com nota média até 2 nos últimos 3 meses.

Fonte: Elaborado pelo autor.

Nota-se que as três principais variáveis que influenciam positivamente são: dias_desde_ult_venda_Mais_90, qtd_pedidos_cancelados_ultimo_mes_Mais_de_1 e qtd_pedidos_outros_status_ultimo_mes_Mais_de_1. Enquanto isso, tem-se que as três principais variáveis que influenciam negativamente são: qtd_produtos_ultimo_mes_Mais_30, qtd_categorias_vendendo_Mais_60 e qtd_review_ultimos_3_meses_Mais_20.

7.9 Qualidade da previsão

Finalmente foi feita a previsão para a base de teste, a partir desta etapa todos os resultados da qualidade da previsão do modelo são obtidos em uma base de teste que não foi alterada, o objetivo é simular um cenário de previsão futura. Como descobriu-se no contato com a área de negócios que o foco do modelo é ordenação, portanto, a principal métrica que foi analisada é o lift, mas para fins didáticos e de aprendizado foi feita a matriz de confusão, bem como, a curva ROC.

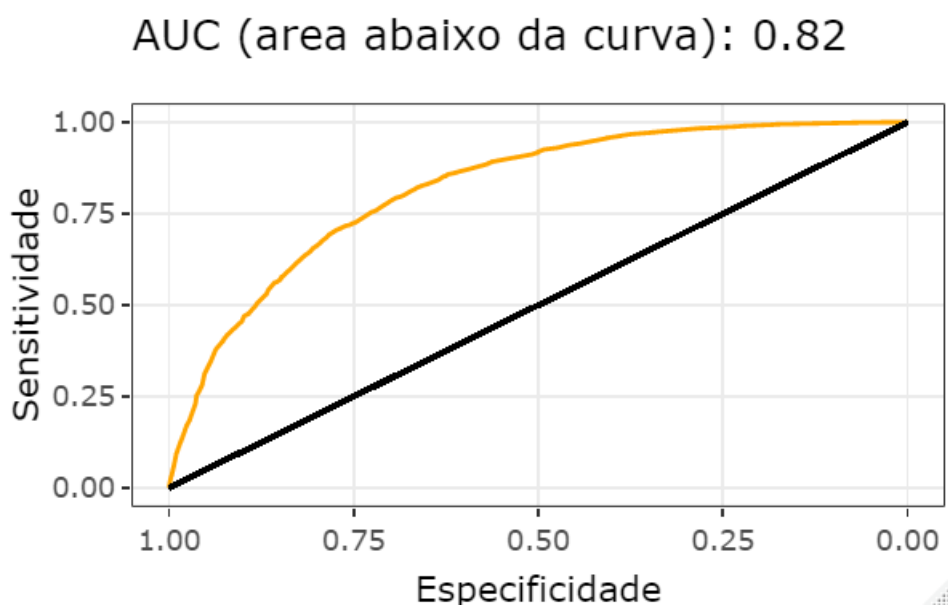
Figura 8 – Matriz de confusão previsão – Base de dados de teste

		Previsto (modelo)	
		0	1
Real	0	3280	370
	1	1166	1025

Fonte: Elaborado pelo autor.

A partir da matriz obteve-se as métricas de acurácia, precisão e recall. Acurácia: 0,737, o que é muito bom pois a acurácia na base de treino era 0,7375. Precisão: 0,7348, é outra medida que acompanha o resultado da base de treino. Recall: 0,4678 é muito bom, embora haja uma diminuição, não é nada muito fora do esperado já que o valor era de 0,4762. Até agora o modelo aparenta ser um bom modelo preditivo.

Gráfico 29 – Curva ROC previsão



Fonte: Elaborado pelo autor.

A curva ROC mostra-se interessante, aparenta ser um modelo bom e a área abaixo da curva (conhecido como AUC) é de 0,82, o que parece ser muito bom, já que o AUC de treino foi 0,821 e quanto mais próximo de 1 melhor.

Finalmente veremos a métrica lift. Para ela foram criados dois cenários, o primeiro cenário em que a área de marketing da olist tem orçamento para abordar 10% da base de vendedores, já o segundo supõe que a área tenha orçamento para abordar 20% dos vendedores.

Para o primeiro cenário tem-se um lift de 2,1044, ou seja, pode-se dizer para a área de marketing que o modelo é 110,44% (ou 1,1044) vezes melhor do que escolher aleatoriamente os indivíduos. Outra forma de explicar essa informação é dizer que utilizando o modelo atual a área de marketing terá 110,44% mais retorno na campanha. Em comparação com o resultado obtido na base de treino, vemos que o modelo está desempenhando muito bem, já que o resultado era de 2,1654.

Já para o segundo cenário tem-se um lift de 2,0428, assim, dizer que o modelo é 104,28% (ou 1,0428) vezes melhor do que escolher aleatoriamente os indivíduos. Também pode-se dizer que a área de marketing terá 104,28% mais retorno na campanha ao utilizar o modelo proposto. Nota-se que neste corte acontece um evento raro que é quando o valor da previsão é melhor que o valor do treino já que o valor anterior era de 2,0011.

O principal ponto a se observar em toda a análise da qualidade da previsão é o quão próxima as medidas da previsão estarão das medidas criadas com os dados de treino e, neste caso, nota-se que as medidas estão muito próximas o que mostra que esse modelo é bom para prever para dados novos, ou seja, ele será um bom modelo para prever o vendedor que não efetuará venda no próximo mês.

Essas informações se explicadas de forma clara e simples têm um grande impacto nas empresas, esse é um dos maiores desafios desta área, trazer o conhecimento teórico e técnico de forma simples e fácil para as pessoas.

8 CONSIDERAÇÕES FINAIS

A criação da ABT foi de longe a tarefa mais trabalhosa e excedeu o tempo previsto no projeto. Criar o código, mesmo que baseado no código do professor Sergio Zaccarelli não foi tarefa fácil, afinal SQL não é uma linguagem tão trivial, principalmente quando se trata de criar tabelas e bases de dados. Como são mais de 200 variáveis, esse código ficou muito extenso.

Ao longo do projeto a passagem dos dados do SQL para o R foi um problema, eles ficavam sobrescritos e a cada tentativa a base aumentava cada vez mais, graças ao Alberto Surano, colega do curso da ASN, esse problema foi resolvido.

A análise descritiva dos dados é uma tarefa muito importante. Dado o grande número de variáveis, a análise descritiva ficou muito extensa, portanto, visando encurtar a análise e focar nas ideias mais importantes, apenas os principais resultados da análise descritiva foram mostrados no relatório final.

Após a descritiva houve a criação do modelo, mas o principal desafio foi como apresentar os resultados de forma simples e clara neste relatório. É muito difícil apresentar formalmente e de forma escrita esses resultados, mas acredita-se que esse objetivo foi atingido.

O modelo de Regressão Logística funcionou bem e fez previsões consistentes, ou seja, é um bom modelo para previsões futuras. Com isso, a empresa Olist poderia utilizá-lo para selecionar os vendedores que não farão venda no mês seguinte e a partir dessa seleção, oferecer uma ajuda personalizada para que esses vendedores tenham maior probabilidade de venderem no mês seguinte.

A principal referência do trabalho é o livro de Hosmer e Lemeshow (2013) que é o livro mais completo sobre regressão logística que foi encontrado.

Referências

- CABRAL, C. I. S. **Aplicação do Modelo de Regressão Logística num Estudo de Mercado**. Dissertação (Mestrado em Matemática Aplicada à Economia e à Gestão) - Universidade de Lisboa Faculdade de Ciências, Lisboa, 2013.
- BATISTA, J. L. F. **Verossimilhança e Máxima Verossimilhança**. Esalq-USP (2009). Disponível em: <http://cmq.esalq.usp.br/BIE5781/lib/exe/fetch.php?media=leituras:verossim.pdf>> Acesso em: 01/07/2022.
- HOSMER, D.W.; LEMESHOW, S. **Applied logistic regression**. Inglês (Estados Unidos) 3rd ed. New Jersey: John Wiley & Sons, 2013. 500 p.
- METHA, N. STEINMAN, D. MURPHY, L. MARTINEZ, M. **Customer Success: How Innovate Companies Are Reducing Churn and Growing Recurring Revenue**. New York: Wiley, 2016. 265 p.
- PAULA, G. A. **Modelos de regressão com apoio computacional**. Instituto de matemática e Estatística. Universidade de São Paulo. Disponível em: https://www.ime.usp.br/~qiapaula/texto_2013.pdf>. Acesso em 20/09/2021.
- PROVOST, F.; FAWCETT, T. **Data Science para negócios: O que você precisa saber sobre mineração de dados e pensamento analítico de dados**. Rio de Janeiro: Alta books, 1ªed. 404p. 2016.
- SAS. **ETL: o que é e qual sua importância?** SAS. Disponível em: https://www.sas.com/pt_br/insights/data-management/o-que-e-etl.html>. Acesso em: 16/01/2022
- SILVA, A. M. M. **Notas de aula do curso Data Science analytics: Uma visão geral**. (Apostila). 2020. São Paulo: [s.n.]. Apostila disponível apenas para alunos da empresa ASN.Rocks.
- SILVA, A. M. M. **Notas de aula do curso Data Science analytics: Transformando Negócios**. (Apostila). 2021. São Paulo: [s.n.]. Apostila disponível apenas para alunos da empresa ASN.Rocks.
- SILVA, A. M. M. **Como utilizar um modelo: 3 Razões detalhadas** ASN.Rocks. 2019. E-book disponível em: <https://asn.rocks/aulas-ebooks/>>. Último acesso: 20/01/2022.
- SILVA, A. M. M. **Métricas de qualidade: Insights para um modelo campeão!** ASN.Rocks. 2019. E-book disponível em: <https://asn.rocks/aulas-ebooks/>>. Acesso em: 20/01/2022.
- SILVA, A. M. M. **Machine Learning: Não cometa os 5 erros clássicos** ASN.Rocks. 2019. E-book disponível em: <https://asn.rocks/aulas-ebooks/>>. Acesso em: 20/01/2022.

WHEELAN, C. **Estatística: O que é, para que serve, como funciona.** Inglês (Estados Unidos). Rio de Janeiro: Zahar, 1ªed. 328p. 2016.

FAWCETT, T. 2006. **An introduction to ROC analysis.** Inglês (Estados Unidos) Pattern Recognition Letters 27: 861–74. Disponível em: <https://doi.org/10.1016/j.irbm.2014.09.001>.

FONSECA, M. **Olist se torna o mais novo unicórnio brasileiro, incluindo empreendedores no e-commerce** Infomoney (2021). Disponível em: <https://www.infomoney.com.br/do-zero-ao-topo/olist-se-torna-mais-novo-unicornio-brasileiro-incluindo-empresendedores-no-e-commerce/>> Acesso em: 07/03/2022.

FONSECA, M. **Olist: o unicórnio do e-commerce que nasceu de uma loja de artesanato** Infomoney (2022). Disponível em: <https://www.infomoney.com.br/do-zero-ao-topo/olist-o-unicornio-do-e-commerce-que-nasceu-de-uma-loja-de-artesanato/>> Acesso em: 17/07/2022.

HAIR, JOSEPH F., WILLIAM C. B., BARRY J. B., e RONALD L. T. **Análise Multivariada de Dados.** 2009. 6a ed. São Paulo: Bookman.

OLIVEIRA, C. **Investimento de US\$ 80 milhões no Olist: como o aporte está apoiando nosso crescimento** Olist (2021). Disponível em: <https://olist.com/blog/pt/olist/olist-news/investimento-olist-80-milhoes/>> Acesso em: 08/03/2022.

APÊNDICES

Apêndice 1 –Dicionários dos dados obtidos na plataforma Keaggle

Quadro 1 - Dicionário da base olist_customers_dataset

Nome da variável	Descrição
customer_id	ID do cliente.
customer_code_prefix	Prefixo do código postal (começo do CEP).
customer_city	Cidade do cliente
customer_state	Estado do cliente

Fonte: Elaborado pelo autor.

Quadro 2 - Dicionário da base olist_geolocation_dataset

Nome da variável	Descrição
geolocation_zip_code_prefix	Prefixo do código postal (começo do CEP)
geolocation_lat	Latitude geográfica do cliente.
geolocation_lng	Longitude geográfica
geolocation_city	Cidade do cliente, assim como na base customers.
geolocation_state	Estado do cliente, assim como na base customers.

Fonte: Elaborado pelo autor.

Quadro 3 - Dicionário da base olist_order_items_dataset

Nome da variável	Descrição
order_id	ID do pedido.
order_item_id	Posição do item no pedido, por exemplo, se for comprado um lápis, uma borracha e uma caneta no mesmo pedido, vão aparecer 3 linhas na tabela, a linha onde o produto é o lápis vai aparecer com o order_item_id igual a 1, já o da borracha vai ser igual a 2 e o da caneta igual a 3.
product_id	ID do produto.

seller_id	ID do vendedor
price	Preço do produto
freight	Frete do produto

Fonte: Elaborado pelo autor.

Quadro 4 - Dicionário da base olist_order_payments_dataset

Nome da variável	Descrição
order_id	ID do pedido
payment_sequential	Ordem do pagamento
payment_type	Como foi pago o pedido
payment_installments	Número de parcelas da compra
payment_value	Valor da compra

Fonte: Elaborado pelo autor.

Quadro 5 - Dicionário da base olist_order_reviews_dataset

Nome da variável	Descrição
review_id	ID da avaliação do pedido
order_id	ID do pedido
review_score	Nota dada pelo cliente para aquele pedido
review_comment_title	Título do comentário do cliente
review_comment_message	Mensagem enviada pelo cliente
review_creation_date	Data de criação da avaliação
review_answer_timestamp	Data e hora da resposta da avaliação

Fonte: Elaborado pelo autor.

Quadro 6 - Dicionário da base olist_orders_dataset

Nome da variável	Descrição
order_id	ID do pedido
customer_id	ID do cliente
order_status	Status do pedido
order_purchase_timestamp	Data e hora da compra
order_approved_at	Data e hora da aprovação da compra
order_delivered_carrier_date	Data e hora em que o pedido foi entregue para a transportadora

order_delivered_customer_date	Data e hora em que o pedido foi entregue ao cliente
order_estimated_delivery_date	Data de entrega estimada (previsão de entrega)

Fonte: Elaborado pelo autor.

Quadro 7 - Dicionário da base olist_products_dataset

Nome da variável	Descrição
product_id	ID do produto
product_category_name	Nome da categoria a qual o produto pertence
product_name_lenght	Número de caracteres no nome do produto
product_description_lenght	Número de caracteres na descrição do produto
product_photos_qty	Quantidade de fotos do produto
product_weight_g	Peso do produto em gramas(g)
product_lenght_cm	Comprimento do produto em centímetros(cm)
product_height_cm	Altura do produto em centímetros (cm)
product_width_cm	Largura do produto em centímetros (cm)

Fonte: Elaborado pelo autor.

Quadro 8 - Dicionário da base olist_sellers_dataset

Nome da variável	Descrição
seller_id	ID do vendedor
seller_zip_code_prefix	Prefixo do código postal (começo do CEP) do vendedor
seller_city	Cidade do vendedor
seller_state	Estado do vendedor

Apêndice 2 –Dicionário dos dados da ABT

Quadro 9 - Dicionário da ABT

Nome da variável	Descrição
sellerId	Identificação/código do vendedor
fl_venda	(1=se teve venda no período de 1 mês 0 = se não teve venda no período de 1 mês)
cep_vendedor	prefix do CEP do vendedor
cidade_vendedor	Cidade do Vendedor
estado_vendedor	Estado do Vendedor
latitude_vendedor	Posição latitudinal do vendedor
longitude_vendedor	Posição longitudinal do vendedor
dias_desde_ult_venda	Dias desde a ultima compra
dias_desde_ult_estimativa_entrega	Dias desde a ultima estimativa de entrega
dias_desde_ult_entrega_pedido	Dias desde a última entrega ao cliente
qtd_dias_ultima_venda	Quantidade de dias desde a última venda
qtd_pedidos_ate_hoje	Quantidade de pedidos até hoje
qtd_pedidos_ultimo_mes	Quantidade de pedidos no último mês
qtd_pedidos_ultimos_2_meses	Quantidade de pedidos nos últimos 2 meses
qtd_pedidos_ultimos_3_meses	Quantidade de pedidos nos últimos 3 meses
qtd_pedidos_ultimos_4_meses	Quantidade de pedidos nos últimos 4 meses
qtd_pedidos_ultimos_5_meses	Quantidade de pedidos nos últimos 5 meses
qtd_pedidos_ultimos_6_meses	Quantidade de pedidos nos últimos 6 meses
tend_pedidos_m1_m3	Tendência média da quantidade de pedidos no último mês em relação aos últimos 3 meses.
tend_pedidos_m1_m6	Tendência média da quantidade de pedidos no último mês em relação aos últimos 6 meses.
tend_pedidos_m3_m6	Tendência média da quantidade de pedidos nos último 3 meses em relação aos últimos 6 meses.
qtd_produtos_ate_hoje	Quantidade de produtos até hoje
qtd_produtos_ultimo_mes	Quantidade de produtos no último mês
qtd_produtos_ultimos_2_meses	Quantidade de produtos nos últimos 2 meses
qtd_produtos_ultimos_3_meses	Quantidade de produtos nos últimos 3 meses

qtd_produtos_ultimos_4_meses	Quantidade de produtos nos últimos 4 meses
qtd_produtos_ultimos_5_meses	Quantidade de produtos nos últimos 5 meses
qtd_produtos_ultimos_6_meses	Quantidade de produtos nos últimos 6 meses
media_dias_entre_dt_compra_dt_entrega_ate_hoje	Média de dias entre a data da compra e a data estimada da entrega até hoje
media_dias_entre_dt_compra_dt_entrega_ultimo_mes	Média de dias entre a data da compra e a data estimada da entrega no último mês
media_dias_entre_dt_compra_dt_entrega_ultimos_2_meses	Média de dias entre a data da compra e a data estimada da entrega nos últimos 2 meses
media_dias_entre_dt_compra_dt_entrega_ultimos_3_meses	Média de dias entre a data da compra e a data estimada da entrega nos últimos 3 meses
media_dias_entre_dt_compra_dt_entrega_ultimos_4_meses	Média de dias entre a data da compra e a data estimada da entrega nos últimos 4 meses
media_dias_entre_dt_compra_dt_entrega_ultimos_5_meses	Média de dias entre a data da compra e a data estimada da entrega nos últimos 5 meses
media_dias_entre_dt_compra_dt_entrega_ultimos_6_meses	Média de dias entre a data da compra e a data estimada da entrega nos últimos 6 meses
media_dias_entre_dt_compra_dt_entrega_ate_hoje	Média de dias entre a data da compra e a data de entrega até hoje
media_dias_entre_dt_compra_dt_entrega_ultimo_mes	Média de dias entre a data da compra e a data de entrega no último mês
media_dias_entre_dt_compra_dt_entrega_ultimos_2_meses	Média de dias entre a data da compra e a data de entrega nos últimos 2 meses
media_dias_entre_dt_compra_dt_entrega_ultimos_3_meses	Média de dias entre a data da compra e a data de entrega nos últimos 3 meses
media_dias_entre_dt_compra_dt_entrega_ultimos_4_meses	Média de dias entre a data da compra e a data de entrega nos últimos 4 meses
media_dias_entre_dt_compra_dt_entrega_ultimos_5_meses	Média de dias entre a data da compra e a data de entrega nos últimos 5 meses
media_dias_entre_dt_compra_dt_entrega_ultimos_6_meses	Média de dias entre a data da compra e a data de entrega nos últimos 6 meses
media_dias_entre_dt_estimada_dt_entrega_ate_hoje	Média de dias entre a data estimada e a data de entrega até hoje

media_dias_entre_dt_estimada_dt_entrega_ultimo_mes	Média de dias entre a data estimada e a data de entrega no último mês
media_dias_entre_dt_estimada_dt_entrega_ultimos_2_meses	Média de dias entre a data estimada e a data de entrega nos últimos 2 meses
media_dias_entre_dt_estimada_dt_entrega_ultimos_3_meses	Média de dias entre a data estimada e a data de entrega nos últimos 3 meses
media_dias_entre_dt_estimada_dt_entrega_ultimos_4_meses	Média de dias entre a data estimada e a data de entrega nos últimos 4 meses
media_dias_entre_dt_estimada_dt_entrega_ultimos_5_meses	Média de dias entre a data estimada e a data de entrega nos últimos 5 meses
media_dias_entre_dt_estimada_dt_entrega_ultimos_6_meses	Média de dias entre a data estimada e a data de entrega nos últimos 6 meses
media_vlr_pedido_ate_hoje	Média do valor dos pedidos até hoje
media_vlr_pedido_ultimo_mes	Média do valor dos pedidos no último mês
media_vlr_pedido_ultimos_2_meses	Média do valor dos pedidos nos últimos 2 meses
media_vlr_pedido_ultimos_3_meses	Média do valor dos pedidos nos últimos 3 meses
media_vlr_pedido_ultimos_4_meses	Média do valor dos pedidos nos últimos 4 meses
media_vlr_pedido_ultimos_5_meses	Média do valor dos pedidos nos últimos 5 meses
media_vlr_pedido_ultimos_6_meses	Média do valor dos pedidos nos últimos 6 meses
tend_media_vlr_pedido_m1_m3	Tendência média do valor médio dos pedidos no último mês em relação aos últimos 3 meses.
tend_media_vlr_pedido_m1_m6	Tendência média do valor médio dos pedidos no último mês em relação aos últimos 6 meses.
tend_media_vlr_pedido_m3_m6	Tendência média do valor médio dos pedidos nos último 3 meses em relação aos últimos 6 meses.
media_vlr_frete_ate_hoje	Média do valor dos pedidos até hoje
media_vlr_frete_ultimo_mes	Média do valor dos pedidos no último mês
media_vlr_frete_ultimos_2_meses	Média do valor dos pedidos nos últimos 2 meses
media_vlr_frete_ultimos_3_meses	Média do valor dos pedidos nos últimos 3 meses
media_vlr_frete_ultimos_4_meses	Média do valor dos pedidos nos últimos 4 meses

media_vlr_frete_ultimos_5_meses	Média do valor dos pedidos nos últimos 5 meses
media_vlr_frete_ultimos_6_meses	Média do valor dos pedidos nos últimos 6 meses
media_do_perc_valor_frete_contra_produto	Média do valor percentual referente ao frete em relação ao valor do produto
media_do_perc_valor_frete_contra_produto_ultimo_mes	Média do valor percentual referente ao frete em relação ao valor do produto no último mês
media_do_perc_valor_frete_contra_produto_ultimos_2_meses	Média do valor percentual referente ao frete em relação ao valor do produto nos últimos 2 meses
media_do_perc_valor_frete_contra_produto_ultimos_3_meses	Média do valor percentual referente ao frete em relação ao valor do produto nos últimos 3 meses
media_do_perc_valor_frete_contra_produto_ultimos_4_meses	Média do valor percentual referente ao frete em relação ao valor do produto nos últimos 4 meses
media_do_perc_valor_frete_contra_produto_ultimos_5_meses	Média do valor percentual referente ao frete em relação ao valor do produto nos últimos 5 meses
media_do_perc_valor_frete_contra_produto_ultimos_6_meses	Média do valor percentual referente ao frete em relação ao valor do produto nos últimos 6 meses
qtd_pedidos_entregues_ate_hoje	Quantidade pedidos entregues até hoje
qtd_pedidos_entregues_ultimo_mes	Quantidade de pedidos entregues no último mês
qtd_pedidos_entregues_ultimos_2_meses	Quantidade de pedidos entregues nos últimos 2 meses
qtd_pedidos_entregues_ultimos_3_meses	Quantidade de pedidos entregues nos últimos 3 meses
qtd_pedidos_entregues_ultimos_4_meses	Quantidade de pedidos entregues nos últimos 4 meses
qtd_pedidos_entregues_ultimos_5_meses	Quantidade de pedidos entregues nos últimos 5 meses
qtd_pedidos_entregues_ultimos_6_meses	Quantidade de pedidos entregues nos últimos 6 meses
qtd_pedidos_cancelados_ate_hoje	Quantidade pedidos entregues até hoje
qtd_pedidos_cancelados_ultimo_mes	Quantidade de pedidos entregues no último mês
qtd_pedidos_cancelados_ultimos_2_meses	Quantidade de pedidos entregues nos últimos 2 meses
qtd_pedidos_cancelados_ultimos_3_meses	Quantidade de pedidos entregues nos últimos 3 meses
qtd_pedidos_cancelados_ultimos_4_meses	Quantidade de pedidos entregues nos últimos 4 meses

qtd_pedidos_cancelados_ultimos_5_meses	Quantidade de pedidos entregues nos últimos 5 meses
qtd_pedidos_cancelados_ultimos_6_meses	Quantidade de pedidos entregues nos últimos 6 meses
qtd_pedidos_outros_status_ate_hoje	Quantidade pedidos outros status até hoje
qtd_pedidos_outros_status_ultimo_mes	Quantidade de pedidos outros status no último mês
qtd_pedidos_outros_status_ultimos_2_meses	Quantidade de pedidos outros status nos últimos 2 meses
qtd_pedidos_outros_status_ultimos_3_meses	Quantidade de pedidos outros status nos últimos 3 meses
qtd_pedidos_outros_status_ultimos_4_meses	Quantidade de pedidos outros status nos últimos 4 meses
qtd_pedidos_outros_status_ultimos_5_meses	Quantidade de pedidos outros status nos últimos 5 meses
qtd_pedidos_outros_status_ultimos_6_meses	Quantidade de pedidos outros status nos últimos 6 meses
media_valores_pedidos_entregues_ate_hoje	Média do valor dos pedidos pedidos entregues até hoje
media_valores_pedidos_entregues_ultimo_mes	Média do valor dos pedidos pedidos entregues no último mês
media_valores_pedidos_entregues_ultimos_2_meses	Média do valor dos pedidos pedidos entregues nos últimos 2 meses
media_valores_pedidos_entregues_ultimos_3_meses	Média do valor dos pedidos pedidos entregues nos últimos 3 meses
media_valores_pedidos_entregues_ultimos_4_meses	Média do valor dos pedidos pedidos entregues nos últimos 4 meses
media_valores_pedidos_entregues_ultimos_5_meses	Média do valor dos pedidos pedidos entregues nos últimos 5 meses
media_valores_pedidos_entregues_ultimos_6_meses	Média do valor dos pedidos pedidos entregues nos últimos 6 meses
media_valores_pedidos_cancelados_ate_hoje	Média do valor dos pedidos pedidos cancelados até hoje
media_valores_pedidos_cancelados_ultimo_mes	Média do valor dos pedidos pedidos cancelados no último mês
media_valores_pedidos_cancelados_ultimos_2_meses	Média do valor dos pedidos pedidos cancelados nos últimos 2 meses
media_valores_pedidos_cancelados_ultimos_3_meses	Média do valor dos pedidos pedidos cancelados nos últimos 3 meses
media_valores_pedidos_cancelados_ultimos_4_meses	Média do valor dos pedidos pedidos cancelados nos últimos 4 meses
media_valores_pedidos_cancelados_ultimos_5_meses	Média do valor dos pedidos pedidos cancelados nos últimos 5 meses
media_valores_pedidos_cancelados_ultimos_6_meses	Média do valor dos pedidos pedidos cancelados nos últimos 6 meses
qtd_pedidos_dt_prevista_menor_data_entrega	Quantidade de pedidos com Data Prevista de Entrega menor que Data de Entrega até hoje

qtd_pedidos_dt_prevista_menor_data_entrega_ultimo_mes	Quantidade de pedidos com Data Prevista de Entrega menor que Data de Entrega no último mês
qtd_pedidos_dt_prevista_menor_data_entrega_ultimos_2_meses	Quantidade de pedidos com Data Prevista de Entrega menor que Data de Entrega nos últimos 2 meses
qtd_pedidos_dt_prevista_menor_data_entrega_ultimos_3_meses	Quantidade de pedidos com Data Prevista de Entrega menor que Data de Entrega nos últimos 3 meses
qtd_pedidos_dt_prevista_menor_data_entrega_ultimos_4_meses	Quantidade de pedidos com Data Prevista de Entrega menor que Data de Entrega nos últimos 4 meses
qtd_pedidos_dt_prevista_menor_data_entrega_ultimos_5_meses	Quantidade de pedidos com Data Prevista de Entrega menor que Data de Entrega nos últimos 5 meses
qtd_pedidos_dt_prevista_menor_data_entrega_ultimos_6_meses	Quantidade de pedidos com Data Prevista de Entrega menor que Data de Entrega nos últimos 6 meses
qtd_pedidos_pgto_cc_ultimo_mes	Quantidade de pedidos com pagamento cartao de credito no último mês
qtd_pedidos_pgto_cc_ultimos_2_meses	Quantidade de pedidos com pagamento cartao de credito nos últimos 2 meses
qtd_pedidos_pgto_cc_ultimos_3_meses	Quantidade de pedidos com pagamento cartao de credito nos últimos 3 meses
qtd_pedidos_pgto_cc_ultimos_4_meses	Quantidade de pedidos com pagamento cartao de credito nos últimos 4 meses
qtd_pedidos_pgto_cc_ultimos_5_meses	Quantidade de pedidos com pagamento cartao de credito nos últimos 5 meses
qtd_pedidos_pgto_cc_ultimos_6_meses	Quantidade de pedidos com pagamento cartao de credito nos últimos 6 meses
qtd_pedidos_pgto_boleto_ultimo_mes	Quantidade de pedidos com pagamento boleto no último mês
qtd_pedidos_pgto_boleto_ultimos_2_meses	Quantidade de pedidos com pagamento boleto nos últimos 2 meses
qtd_pedidos_pgto_boleto_ultimos_3_meses	Quantidade de pedidos com pagamento boleto nos últimos 3 meses
qtd_pedidos_pgto_boleto_ultimos_4_meses	Quantidade de pedidos com pagamento boleto nos últimos 4 meses
qtd_pedidos_pgto_boleto_ultimos_5_meses	Quantidade de pedidos com pagamento boleto nos últimos 5 meses

qtd_pedidos_pgto_boleto_ultimos_6_meses	Quantidade de pedidos com pagamento boleto nos últimos 6 meses
qtd_pedidos_pgto_outros_ultimo_mes	Quantidade de pedidos com pagamento outros (não cartao, não boleto) no último mês
qtd_pedidos_pgto_outros_ultimos_2_meses	Quantidade de pedidos com pagamento outros (não cartao, não boleto) nos últimos 2 meses
qtd_pedidos_pgto_outros_ultimos_3_meses	Quantidade de pedidos com pagamento outros (não cartao, não boleto) nos últimos 3 meses
media_qtd_parcelas_pedidos_pgto_cc_ultimo_mes	Média da quantidade de parcelas dos pedidos com pagamento cartao de credito no último mês
media_qtd_parcelas_pedidos_pgto_cc_ultimos_2_meses	Média da quantidade de parcelas dos pedidos com pagamento cartao de credito nos últimos 2 meses
media_qtd_parcelas_pedidos_pgto_cc_ultimos_3_meses	Média da quantidade de parcelas dos pedidos com pagamento cartao de credito nos últimos 3 meses
media_qtd_parcelas_pedidos_pgto_cc_ultimos_4_meses	Média da quantidade de parcelas dos pedidos com pagamento cartao de credito nos últimos 4 meses
media_qtd_parcelas_pedidos_pgto_cc_ultimos_5_meses	Média da quantidade de parcelas dos pedidos com pagamento cartao de credito nos últimos 5 meses
media_qtd_parcelas_pedidos_pgto_cc_ultimos_6_meses	Média da quantidade de parcelas dos pedidos com pagamento cartao de credito nos últimos 6 meses
media_qtd_parcelas_pedidos_pgto_boleto_ultimo_mes	Média da quantidade de parcelas dos pedidos com pagamento boleto no último mês
media_qtd_parcelas_pedidos_pgto_boleto_ultimos_2_meses	Média da quantidade de parcelas dos pedidos com pagamento boleto nos últimos 2 meses
media_qtd_parcelas_pedidos_pgto_boleto_ultimos_3_meses	Média da quantidade de parcelas dos pedidos com pagamento boleto nos últimos 3 meses
media_qtd_parcelas_pedidos_pgto_boleto_ultimos_4_meses	Média da quantidade de parcelas dos pedidos com pagamento boleto nos últimos 4 meses
media_qtd_parcelas_pedidos_pgto_boleto_ultimos_5_meses	Média da quantidade de parcelas dos pedidos com pagamento boleto nos últimos 5 meses
media_qtd_parcelas_pedidos_pgto_boleto_ultimos_6_meses	Média da quantidade de parcelas dos pedidos com pagamento boleto nos últimos 6 meses

media_qtd_parcelas_pedidos_pgto_outros_ultimo_mes	Média da quantidade de parcelas dos pedidos com pagamento outros (não cartao, não boleto) no último mês
media_qtd_parcelas_pedidos_pgto_outros_ultimos_2_meses	Média da quantidade de parcelas dos pedidos com pagamento outros (não cartao, não boleto) nos últimos 2 meses
media_qtd_parcelas_pedidos_pgto_outros_ultimos_3_meses	Média da quantidade de parcelas dos pedidos com pagamento outros (não cartao, não boleto) nos últimos 3 meses
media_qtd_parcelas_pedidos_pgto_outros_ultimos_4_meses	Média da quantidade de parcelas dos pedidos com pagamento outros (não cartao, não boleto) nos últimos 4 meses
media_qtd_parcelas_pedidos_pgto_outros_ultimos_5_meses	Média da quantidade de parcelas dos pedidos com pagamento outros (não cartao, não boleto) nos últimos 5 meses
media_qtd_parcelas_pedidos_pgto_outros_ultimos_6_meses	Média da quantidade de parcelas dos pedidos com pagamento outros (não cartao, não boleto) nos últimos 6 meses
max_qtd_parcelas_pedidos_pgto_cc_ultimo_mes	Máximo da quantidade de parcelas dos pedidos com pagamento cartao de credito no último mês
max_qtd_parcelas_pedidos_pgto_cc_ultimos_2_meses	Máximo da quantidade de parcelas dos pedidos com pagamento cartao de credito nos últimos 2 meses
max_qtd_parcelas_pedidos_pgto_cc_ultimos_3_meses	Máximo da quantidade de parcelas dos pedidos com pagamento cartao de credito nos últimos 3 meses
max_qtd_parcelas_pedidos_pgto_cc_ultimos_4_meses	Máximo da quantidade de parcelas dos pedidos com pagamento cartao de credito nos últimos 4 meses
max_qtd_parcelas_pedidos_pgto_cc_ultimos_5_meses	Máximo da quantidade de parcelas dos pedidos com pagamento cartao de credito nos últimos 5 meses
max_qtd_parcelas_pedidos_pgto_cc_ultimos_6_meses	Máximo da quantidade de parcelas dos pedidos com pagamento cartao de credito nos últimos 6 meses
max_qtd_parcelas_pedidos_pgto_boleto_ultimo_mes	Máximo da quantidade de parcelas dos pedidos com pagamento boleto no último mês
max_qtd_parcelas_pedidos_pgto_boleto_ultimos_2_meses	Máximo da quantidade de parcelas dos pedidos com pagamento boleto nos últimos 2 meses

max_qtd_parcelas_pedidos_pgto_boleto_ultimos_3_meses	Máximo da quantidade de parcelas dos pedidos com pagamento boleto nos últimos 3 meses
max_qtd_parcelas_pedidos_pgto_o_boleto_ultimos_4_meses	Máximo da quantidade de parcelas dos pedidos com pagamento boleto nos últimos 4 meses
max_qtd_parcelas_pedidos_pgto_boleto_ultimos_5_meses	Máximo da quantidade de parcelas dos pedidos com pagamento boleto nos últimos 5 meses
max_qtd_parcelas_pedidos_pgto_boleto_ultimos_6_meses	Máximo da quantidade de parcelas dos pedidos com pagamento boleto nos últimos 6 meses
max_qtd_parcelas_pedidos_pgto_outros_ultimo_mes	Máximo da quantidade de parcelas dos pedidos com pagamento outros (não cartao, não boleto) no último mês
max_qtd_parcelas_pedidos_pgto_outros_ultimos_2_meses	Máximo da quantidade de parcelas dos pedidos com pagamento outros (não cartao, não boleto) nos últimos 2 meses
max_qtd_parcelas_pedidos_pgto_outros_ultimos_3_meses	Máximo da quantidade de parcelas dos pedidos com pagamento outros (não cartao, não boleto) nos últimos 3 meses
max_qtd_parcelas_pedidos_pgto_outros_ultimos_4_meses	Máximo da quantidade de parcelas dos pedidos com pagamento outros (não cartao, não boleto) nos últimos 4 meses
max_qtd_parcelas_pedidos_pgto_outros_ultimos_5_meses	Máximo da quantidade de parcelas dos pedidos com pagamento outros (não cartao, não boleto) nos últimos 5 meses
max_qtd_parcelas_pedidos_pgto_outros_ultimos_6_meses	Máximo da quantidade de parcelas dos pedidos com pagamento outros (não cartao, não boleto) nos últimos 6 meses
media_valor_pedidos_pgto_cc_ultimo_mes	Média do valor dos pedidos com pagamento cartao de credito no último mês
media_valor_pedidos_pgto_cc_ultimos_2_meses	Média do valor dos pedidos com pagamento cartao de credito nos últimos 2 meses
media_valor_pedidos_pgto_cc_ultimos_3_meses	Média do valor dos pedidos com pagamento cartao de credito nos últimos 3 meses
media_valor_pedidos_pgto_cc_ultimos_4_meses	Média do valor dos pedidos com pagamento cartao de credito nos últimos 4 meses

media_valor_pedidos_pgto _cc_ultimos_5_meses	Média do valor dos pedidos com pagamento cartao de credito nos últimos 5 meses
media_valor_pedidos_pgto _cc_ultimos_6_meses	Média do valor dos pedidos com pagamento cartao de credito nos últimos 6 meses
media_valor_pedidos_pgto _boleto_ultimo_mes	Média do valor dos pedidos com pagamento boleto no último mês
media_valor_pedidos_pgto _boleto_ultimos_2_meses	Média do valor dos pedidos com pagamento boleto nos últimos 2 meses
media_valor_pedidos_pgto _boleto_ultimos_3_meses	Média do valor dos pedidos com pagamento boleto nos últimos 3 meses
media_valor_pedidos_pgto _boleto_ultimos_4_meses	Média do valor dos pedidos com pagamento boleto nos últimos 4 meses
media_valor_pedidos_pgto _boleto_ultimos_5_meses	Média do valor dos pedidos com pagamento boleto nos últimos 5 meses
media_valor_pedidos_pgto _boleto_ultimos_6_meses	Média do valor dos pedidos com pagamento boleto nos últimos 6 meses
media_valor_pedidos_pgto _outros_ultimo_mes	Média do valor dos pedidos com pagamento outros (não cartao, não boleto) no último mês
media_valor_pedidos_pgto _outros_ultimos_2_meses	Média do valor dos pedidos com pagamento outros (não cartao, não boleto) nos últimos 2 meses
media_valor_pedidos_pgto _outros_ultimos_3_meses	Média do valor dos pedidos com pagamento outros (não cartao, não boleto) nos últimos 3 meses
media_valor_pedidos_pgto _outros_ultimos_4_meses	Média do valor dos pedidos com pagamento outros (não cartao, não boleto) nos últimos 4 meses
media_valor_pedidos_pgto _outros_ultimos_5_meses	Média do valor dos pedidos com pagamento outros (não cartao, não boleto) nos últimos 5 meses
media_valor_pedidos_pgto _outros_ultimos_6_meses	Média do valor dos pedidos com pagamento outros (não cartao, não boleto) nos últimos 6 meses
quantidade_de_produtos _por_categoria	Número médio de produtos que vende por categoria de venda
media_num_fotos_por _produto	Número médio de fotos por produto que vende
media_fotos_por produto_categoria	Número médio de fotos por categoria
media_tamanho_descricao _por_produto	Tamanho médio do length de cada produto
qtd_categorias_vendendo	Quantidade de categorias que atua

qtd_reviews_ate_hoje	Quantidade de Reviews até hoje
media_notas_review_ate_hoje	Nota média recebida nas Reviews até hoje
media_dias_entre_criacao_resposta_review_ate_hoje	Média da quantidade de dias na resposta do Review recebidos até hoje
qtd_review_ultimo_mes	Quantidade Recebida de Reviews no último mês
qtd_review_ultimos_2_meses	Quantidade Recebida de Reviews nos últimos 2 meses
qtd_review_ultimos_3_meses	Quantidade Recebida de Reviews nos últimos 3 meses
qtd_review_ultimos_4_meses	Quantidade Recebida de Reviews nos últimos 4 meses
qtd_review_ultimos_5_meses	Quantidade Recebida de Reviews nos últimos 5 meses
qtd_review_ultimos_6_meses	Quantidade Recebida de Reviews nos últimos 6 meses
tend_qtd_notas_review_m1_m3	Tendência média da quantidade de notas de review no último mês em relação aos últimos 3 meses.
tend_qtd_notas_review_m1_m6	Tendência média da quantidade de notas de review no último mês em relação aos últimos 6 meses.
tend_qtd_notas_review_m3_m6	Tendência média da quantidade de notas de review nos último 3 meses em relação aos últimos 6 meses.
media_notas_review_ultimo_mes	Nota média recebida nas Reviews do último mês
media_notas_review_ultimos_2_meses	Nota média recebida nas Reviews dos últimos 2 meses
media_notas_review_ultimos_3_meses	Nota média recebida nas Reviews dos últimos 3 meses
media_notas_review_ultimos_4_meses	Nota média recebida nas Reviews dos últimos 4 meses
media_notas_review_ultimos_5_meses	Nota média recebida nas Reviews dos últimos 5 meses
media_notas_review_ultimos_6_meses	Nota média recebida nas Reviews dos últimos 6 meses
tend_notas_review_m1_m3	Tendência média da nota de review no último mês em relação aos últimos 3 meses.
tend_notas_review_m1_m6	Tendência média da nota de review no último mês em relação aos últimos 6 meses.
tend_notas_review_m3_m6	Tendência média da nota de review nos último 3 meses em relação aos últimos 6 meses.

media_dias_entre_criacao_resposta_review_ultimo_mes	Média da quantidade de dias na resposta do Review recebidos no último mês
media_dias_entre_criacao_resposta_review_ultimos_2_meses	Média da quantidade de dias na resposta do Review recebidos nos últimos 2 meses
media_dias_entre_criacao_resposta_review_ultimos_3_meses	Média da quantidade de dias na resposta do Review recebidos nos últimos 3 meses
media_dias_entre_criacao_resposta_review_ultimos_4_meses	Média da quantidade de dias na resposta do Review recebidos nos últimos 4 meses
media_dias_entre_criacao_resposta_review_ultimos_5_meses	Média da quantidade de dias na resposta do Review recebidos nos últimos 5 meses
media_dias_entre_criacao_resposta_review_ultimos_6_meses	Média da quantidade de dias na resposta do Review recebidos nos últimos 6 meses

Fonte: Elaborado pelo autor.

Apêndice 3 – Link e QR-Code

Link para acesso dos arquivos deste projeto: https://drive.google.com/drive/folders/1K2uoqJMRWps6_nBjJlR4ZL2WL6tcAL2?usp=sharing

QR-Code para acesso dos arquivos deste projeto: <https://qrco.de/bdhV8Z>

