Universidade Federal da Paraíba
Centro de Ciências Exatas e da Natureza
Graduação em Estatística

Comparação de Desempenho e Adequação de três processos para séries temporais INAR

Paulo Manoel da Silva Junior

João Pessoa - PB
2025

Paulo Manoel da Silva Junior

Comparação de Desempenho e Adequação de três processos para séries temporais INAR

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Estatística do Centro de Ciências Exatas e da Natureza da Universidade Federal da Paraíba (UFPB), como requisito para obtenção do grau de Bacharel em Estatística.

Orientador: Profª. Drª. Tatiene Correia de Souza

João Pessoa - PB
2025

Agradecimentos

Em algum momento da minha vida parei de prestar atenção no presente e passei a viver na ansiedade de um futuro melhor. "Preocupado com uma única folha, você não verá a árvore. Preocupado com uma única árvore você não perceberá toda a floresta. Não se preocupe com um único ponto. Veja tudo em sua plenitude sem se forçar."− Takuan Sōhō. Obrigado Eraldo do passado por decidir mudar.

Meu maior agradecimento é pra mainha e painho, Dona Adeilda e Seu Everaldo. Mãe obrigado por todo o carinho infindável que a senhora me proporciona. Pai obrigado por trabalhar tão duro por nós. Obrigado aos dois pelos valores ensinados, isso nenhuma escola ou universidade poderia me ensinar.

Meu sincero obrigado, em especial, ao professor Marcelo por aceitar me orientar nessa etapa final da graduação. Ao professor Tablada e a professora Everlane, pelo suporte durante minha participação na tutoria e na pesquisa científica, projetos esses na qual só agregaram positivamente na minha vida e carreira. Professora Gilmara, obrigado pela paciência, por toda educação e prestatividade como coordenadora desse curso. Por fim, e não menos importante a todos os professores na qual eu pude ter a oportunidade de ser aluno.

Paulin, Eltin, Arthur, Gleyce, nossos dias resolvendo listas e mais listas de exercícios valeram a pena. A troca de conhecimento que tivemos durante essa jornada foi fundamental na formação da pessoa que sou hoje.

Resumo

Este trabalho aborda o uso da aprendizagem de máquina na tentativa de prever o próximo dia de compra de clientes a partir do histórico de compras. Os dados foram obtidos de uma distribuidora da Paraíba que fornece uma vasta variedade de produtos de diferentes departamentos, descartáveis, químicos, papelarias, equipamentos de proteção. Diante do cenário competitivo e da explosão de dados no mercado, a inteligência comercial tornou-se essencial para antecipar desafios e identificar oportunidades. O estudo abrange a coleta e análise do histórico de compras, além da utilização de segmentação de clientes. Diferentes algoritmos, como Redes Neurais Artificiais, Máquinas de Vetores de Suporte e Florestas Aleatórias, são avaliados em busca dos melhores resultados. Os objetivos incluem extrair informações relevantes, realizar análise exploratória, segmentar clientes e avaliação de modelos e técnicas do campo da aprendizagem de máquina. Espera-se que os resultados contribuam para personalizar estratégias do setor comercial.

Palavras-chave: Aprendizagem de máquina, Inteligência de Mercado, Segmentação RFV, Regressão.

Abstract

This study explores the application of machine learning techniques to predict the next purchase day of customers based on their purchase history. The data was obtained from a company distributor in Paraíba, that provides a wide variety of products from different departments, including disposables, chemicals, stationery, and protective equipment. Given the competitive landscape and the data explosion in the market, business intelligence has become crucial for anticipating challenges and identifying opportunities. The research encompasses the collection and analysis of purchase histories, as well as the utilization of customer segmentation. Various algorithms, such as Artificial Neural Networks, Support Vector Machines, and Random Forests, are evaluated to attain optimal results. Objectives include extracting relevant information, conducting exploratory analysis, segmenting customers, and assessing models and techniques within the field of machine learning. The anticipated outcome is to enhance the personalization of commercial strategies.

Keywords: Machine learning, Business Intelligence, RFV Segmentation, Regression.

Lista de tabelas

Lista de ilustrações

Sumário

1 Introdução

As séries temporais com dados inteiros não negativos, conhecidas como **séries de contagem**, ou **pontos de contagem**, são comuns em diversas áreas, como finanças, epidemiologia, seguros e telecomunicações. Exemplos típicos incluem o número de chamadas telefônicas por hora, a quantidade de novos clientes inadimplentes por dia, a quantidade de seguros vendidos a cada semana, acidentes por dia ou casos de doenças por semana. Por sua natureza discreta, essas séries desafiam a aplicação direta de modelos clássicos como os AR, MA, ARMA ou ARIMA, os quais pressupõem variáveis contínuas e normalmente distribuídas (CAMERON; TRIVEDI, 1998).

Como alternativa aos modelos tradicionais, surgiram os chamados **modelos autorregressivos para dados inteiros não negativos**, notadamente os processos **INAR (Integer value autorregresive)**, que respeitam a estrutura discreta da variável ao incorporar um operador de redução. A formulação original do processo INAR(1) foi introduzida inicialmente por Al-Osh e Alzaid (1987) e McKenzie (1985), sendo baseado no operador *thinning binomial* proposto por steutel1979, o qual permite definir processos autorregressivos respeitando a natureza discreta da variável.

Modelos INAR são particularmente atrativos por possibilitarem a incorporação de dependência temporal entre observações, algo que é essencial em séries de contagem. No modelo INAR(1), essa dependência é representada por meio do operador *thinning binomial*, denotado por $\alpha \circ X_{t-1}$, onde $\alpha \in [0, 1]$ representa a taxa de retenção da série. O termo de inovação, denotado por ϵ_t, por outro lado, representa a contribuição probabilistica que é adicionada ao processo a cada instante do tempo. Geralmente, ele é modelado por uma distribuição discreta, sendo a distribuição Poisson a mais utilizada. Essa escolha facilita as análises e resultados, mas a escolha dessa distribuição traz uma limitação que deve ser levada em consideração, pois, ela pressupõe que a média e variância são iguais. Essa pressuposição em muitos casos não é verdadeira, pois, é observado em muitos casos que a variabilidade é maior ou menor do que a média McKenzie (2003).

Nas últimas décadas, os processos INAR têm se mostrado úteis em diversas aplicações práticas. (PEDELI; KARLIS, 2011) propuseram uma versão bivariada do modelo INAR(1) para descrever o número diário de transações financeiras em bolsas de valores, considerando a correlação entre pares de séries de contagem. Em outro contexto, (FOKIANOS; TJøSTHEIM, 2009) utilizaram modelos baseados em INAR para modelar a frequência de chamados de emergência em serviços hospitalares, com foco na dependência temporal e na sobredispersão dos dados. Já (ZHANG; WANG; ZHU, 2019) aplicaram modelos INAR para analisar séries mensais de casos de tuberculose na China, evidenciando a adequação do modelo em contextos epidemiológicos. Tais aplicações reforçam a

relevância da modelagem de séries de contagem com sobredispersão e motivam a utilização de distribuições alternativas à Poisson, tais como a binomial negativa e a geométrica.

1.1 Objetivos

1.1.1 Objetivo Geral

Este trabalho tem como objetivo analisar o desempenho de dois estimadores dos parâmetros do processo Poisson, Binomial Negativa e Geométrica INAR(1), sendo os estimadores mínimos quadrados condicionais e máxima verossimilhança condicional. Como um segundo objetivo desejamos propor e comparar os mesmos dados as três distribuições consideradas, usando os métodos de estimação e depois realizar previsões um passo a frente. Os estudos de desempenho são feitos por meio de simulações de Monte Carlo, que serão realizadas usando a linguagem R (R, Core Team, 2025).

1.1.2 Objetivos Específicos

1. Verificar se os estimadores propostos atendes as propriedades de um bom estimador;
2. Verificar adequação da melhor distribuição de acordo com os dados;
3. Utilizar o processo na predição de séries temporais verdadeiras;
4. Verificar o desempenho do mesmo processo com a variabilidade da distribuição em um mesmo conjunto de dados.

2 Referencial Teórico

2.1 Processo INAR

2.1.1 Operador *thinning* binomial

Uma séria temporal pode ser definida como: Seja $Y_t, t \in Z$ uma série temporal. Dizemos que $\{Y_t, t \in Z\}$ é um processo autoregressivo, AR(1) se satisfaz a equação recursiva

$$Y_t = \alpha Y_{t-1} + \epsilon_t \tag{2.1}$$

onde $\alpha \in R$ e a sequência $\{\epsilon_t, t \in R\}$ chamada de ruído branco, é uma coleção de variáveis aleatórias não correlacionadas, a média e variância dependem da distribuição.

Como em muitos casos $\alpha \in R$, para uma série temporal de valores inteiros não negativos não é admitido esta representação, mesmo que seja considerado uma distribuição de valores inteiros não negativos para ϵ_t, Y_t não será, necessariamente um valor inteiro.

Na literatura uma das formas de obter modelos que garantem que Y_t seja um valor inteiro não negativo é utilizar um operador conveniente para garantir que as variáveis aleatórias assumam valores inteiros não negativos. Utilizaremos o operador *thinning* binomial, proposto por (STEUTEL; HARN, 1979).

Definição 1: Seja Y uma variável aleatória não negativa que assume valores inteiros e seja $\alpha \in [0, 1]$. O operador *thinning* binomial é definido como:

$$\alpha \circ Y = \sum_{i=1}^{Y} N_i, \tag{2.2}$$

em que as variáveis aleatórias $N_i'^s$ são independentes e identicamente distribuídas (i.i.d.), com distribuição Bernoulli de parâmetro α. As variáveis $N_i'^s$ são chamadas de séries de contagem de $\alpha \circ Y$.

2.1.1.1 Propriedades do operador *thinning* binomial

Sejam X e Y variáveis aleatórias não negativas assumindo valores inteiros. Sejam α e β constantes que pertencem aos reais no intervalo $[0, 1]$ e suponha que a série de contagem de $\alpha \circ Y$ é independente da série de contagem de $\beta \circ X$ são independentes de X e de Y. Então:

1. $0 \circ Y = 0$

2. $1 \circ Y = Y$
3. $\alpha \circ (\beta \circ Y) \stackrel{d}{=} (\alpha\beta) \circ Y$
4. $\alpha \circ (Y + X) \stackrel{d}{=} (\alpha \circ Y) + (\beta \circ X)$
5. $E(\alpha \circ Y) = \alpha E(Y)$
6. $Var(\alpha \circ Y) = \alpha^2 Var(Y) + \alpha(1-\alpha)E(Y)$

em que $X \stackrel{d}{=} Y$ significa que X e Y têm a mesma distribuição.

As demonstrações destas propriedades, bem como de outras do operador *thinning* binomial já foram demonstradas e podem ser encontradas em Gomes (2009), Barcelos (2008) e Silva (2005).

2.1.2 Processo INAR(1)

A partir do que foi definido sobre o operador *thinning* binomial, é possível definir o processo autoregressivo de valores inteiros de ordem um, INAR(1), que foi proposto por (AL-OSH; ALZAID, 1987) e (MCKENZIE, 1985).

Definição 2: Um processo estocástico discreto de valores inteiros não negativos $\{Y_t, t \in Z\}$ é dito ser um processo INAR (1), se satisfaz a seguinte equação de recursão

$$Y_t = \alpha \circ Y_{t-1} + \epsilon_t, \tag{2.3}$$

em que $\alpha \in (0,1)$, $\{\epsilon_t, t \in Z\}$ é uma sequência de variáveis aleatórias i.i.d. de valores inteiros não negativos tal que $E(\epsilon_t) = \mu_\epsilon$, $Var(\epsilon_t) = \sigma_\epsilon^2$ e ϵ_t é independente de Y_s para $s < t$.

O processo INAR (1) pode ser interpretado da seguinte forma: Y_t representa a população de habitantes remanescentes em uma região no caso de migração no tempo t. $\alpha \circ Y_{t-1}$ representa os habitantes daquela região que permaneceram naquela região no tempo $t-1$, ϵ_t representa os novos haabitantes que chegaram para morar naquela região no tempo t e $Y_{t-1} - \alpha \circ Y_{t-1}$ a quantidade de habitantes que saíram daquela região.

De acordo com Du e Li (1991) e Latour (1998) se $\alpha < 1$. então o processo INAR(1) é estacionário. Consideraremos aqui apenas quando $\alpha < 1$, caracterizando assim um processo estacionário.

Segue algumas propriedades do processo INAR(1) que segue para adequação de acordo com a distribuição de ϵ_t

2.1.2.1 Propriedades do Processo INAR(1)

Seja $\{Y_t, t \in Z\}$ um processo INAR(1), com $\mu_\epsilon = E(\epsilon_t)$ e $\sigma_\epsilon^2 = Var(\epsilon_t)$. Então, $\{Y_t, t \in Z\}$ satisfaz as seguintes propriedades:

1. $E(Y_t) = \frac{\mu_\epsilon}{1-\alpha}$
2. $Var(Y_t) = \frac{\sigma_\epsilon^2 + \alpha\mu_\epsilon}{1-\alpha^2}$
3. $E(Y_t|Y_{t-1}) = \alpha Y_{t-1} + \mu_\epsilon$
4. $Var(Y_t|Y_{t-1}) = \alpha(1-\alpha)Y_{t-1} + \sigma_\epsilon^2$
5. $Cov(Y_t, Y_{t+j}) = \alpha^j Var(Y_t)$, para $j \in N$
6. $\rho_Y(h) = Corr(Y_t, Y_{t+h}) = \alpha^h$, para $h \in N$

2.1.3 Processo Poisson INAR(1)

Considere $\{Y_t, t \in Z\}$ um processo INAR(1), ou seja, um processo em que Y_t satisfaz a equação abaixo:

$$Y_t = \alpha \circ Y_{t-1} + \epsilon_t \tag{2.4}$$

Quando $\{\epsilon_t, t \in Z\}$ é um conjunto de variáveis aleatórias independentes com distribuição Poisson de parâmetro λ, então $\{Y_t, t \in Z\}$ é chamado de processo Poisson INAR(1). E nesse caso, a média e variância são iguais. Ou seja, $\mu_\epsilon = \sigma_\epsilon^2 = \lambda$

2.1.3.1 Propriedades do Processo Poisson INAR(1)

Seja $\{Y_t, t \in Z\}$ um processo Poisson INAR(1) então, considerando as propriedades do operador *thinning* binomial que foram apresentadas na seção 2.1.2.1 e também que temos um processo estacionário, ou seja, a média e variância permanecem constantes ao longo do tempo. É possível encontrar as seguintes propriedades.

1. $E(Y_t) = \frac{\lambda}{1-\alpha}$
2. $Var(Y_t) = \frac{\lambda}{1-\alpha}$
3. $E(Y_t|Y_{t-1}) = \alpha Y_{t-1} + \lambda$
4. $Var(Y_t|Y_{t-1}) = \alpha(1-\alpha)Y_{t-1} + \lambda$
5. $Cov(Y_t, Y_{t+j}) = \alpha^j Var(Y_t)$, para $j \in N$
6. $\rho_Y(h) = Corr(Y_t, Y_{t+h}) = \alpha^h$, para $h \in N$

2.1.3.2 Estimadores de Mínimos Quadrados Condicionais

Seja $Y_1, Y_2, Y_3, ..., Y_n$ uma amostra do processo Poisson INAR(1) dado na equação (2.3). Estamos interessados em estimar o vetor de parâmetros $\theta = (\alpha, \lambda)$. Sabemos que:

$$E(Y_t|Y_{t-1}) = \alpha Y_{t-1} + \lambda = g(\theta, Y_{t-1})$$

Considere a função,

$$Q_n(\theta) = \sum_{t=2}^{n} [Y_t - g(\theta, Y_{t-1})]^2$$

Os estimadores de MQC de α e λ são os valores de α e λ que minimizam $Q_n(\theta)$. Depois de derivar $Q_n(\theta)$ em relação a α e λ e sendo igualhada as duas derivadas a zero, obtemos que os estimadores de Minímos quadrados condicionais de α e λ são dados por:

$$\widehat{\alpha}_{MQC} = \frac{\sum_{t=2}^{n} Y_t Y_{t-1} - \frac{1}{n-1} \sum_{t=2}^{n} Y_t \sum_{t=2}^{n} Y_{t-1}}{\sum_{t=2}^{n} Y_{t-1}^{n} Y_{t-1}^2 - \frac{1}{n-1} \left(\sum_{t=2}^{n} Y_{t-1}\right)^2} \tag{2.5}$$

$$\widehat{\lambda}_{MQC} = \frac{1}{n-1} \left(\sum_{t=2}^{n} Y_t - \widehat{\alpha}_{MQC} \sum_{t=2}^{n} Y_{t-1} \right) \tag{2.6}$$

2.1.3.3 Estimadores de máxima verossimilhança condicional

Para o estimador de máxima verossimilhança condicional, temos que $Y_t = \alpha \circ Y_{t-1} + \epsilon_t$, em que $\epsilon_t \sim$ Poisson (λ_t)

O operador *thinning binomial*: $\alpha \circ Y = \sum_{K=0}^{y} B_K, \quad B_K \sim$ Bernoulli (α). Logo, $\alpha \circ Y \sim$ Binomial (γ, α)

A função de log-verossimilhança é dada por:

$$l(\alpha, \lambda) = \sum_{t=2}^{T} ln \;\; P(Y_t|Y_{t-1})$$

A probabilidade de Transição: $P(y_t|y_{t-1})$ é:

$$P(Y_t = y_t|Y_{t-1} = y_{t-1}) = P(\alpha \circ Y_{t-1} + \epsilon_t|Y_{t-1} = y_{t-1})$$
$$= P(\alpha \circ Y_{t-1} + \epsilon_t = y_t)$$

Em que $\alpha \circ Y_{t-1} \sim$ Binomial (y_{t-1}, α) e $\epsilon_t \sim$ Poisson (λ), com $X \perp Y$, então,

$$P(X + Y) = \sum_{j=0}^{k} P(X = j, Y = k - j)$$

$$= \sum_{j=0}^{k} P(X = j)P(Y = k - j)$$

Note que, $0 \leq j \leq n$ e $0 \leq j \leq k$. Logo, $0 \leq j \leq min(n.k)$

Logo,

$$P(X + Y = k) = \sum_{j=0}^{min(n,k)} = P(x = j)P(Y = k - j)$$

$$= \sum_{j=0}^{min(n,k)} \binom{n}{j} \alpha^j (1-\alpha)^{n-j} \frac{e^{-\lambda} \lambda^{k-j}}{(k-j)!}$$

Portanto, a função de log-verossimilhança de um processo poisson INAR (1) é dado por:

$$P(Y_t = y_t | Y_{t-1} = y_{t-1}) = \sum_{j=0}^{min(y_{t-1}, y_t)} \binom{y_{t-1}}{j} \alpha^j (1-\alpha)^{y_{t-1}-j} \frac{e^{-\lambda_t} \lambda_t^{y_t - j}}{(y_t - j)!}$$

$$l(\alpha, \lambda) = \sum_{t=2}^{T} ln \left(\sum_{j=0}^{Min(y_t, y_{t-1})} \binom{y_{t-1}}{j} \alpha^j (1-\alpha)^{y_{t-1}-j} \frac{e^{-\lambda t} \lambda t^{y_t - j}}{(y_t - j)!} \right) \tag{2.7}$$

2.1.4 Processo Geométrica INAR(1)

Conforme sugerido por (MCKENZIE, 1986) em processos INAR(1) podem ser utilizados a distribuição **Geométrica e Binomial Negativa**, sendo assim.

Considere $\{Y_t, t \in Z\}$ um processo INAR(1), ou seja, um processo em que Y_t satisfaz a equação (2.3). Quando $\epsilon_t, t \in Z$ é um conjunto de variáveis aleatórias independentes com distribuição Geométrica parametrizada pela média, então $\{Y_t, t \in Z\}$ é chamado de processo Geométrica INAR(1). E nesse caso, temos os seguintes resultados:

Para uma distribuição geométrica com em 0,1,2,3,...

$$P(X = k) = (1-p)^k p, \quad k = 0, 1, 2, 3, ...$$

Parametrizando pela média, temos os seguintes resultados.

- $p = \frac{1}{1+\mu}$
- $P(X = k) = \frac{\mu^k}{(\mu+1)^{k+1}}$
- $E(X) = \mu$
- $Var(X) = \frac{\mu}{(\mu+1)^2}$

2.1.4.1 Propriedades do Processo Geométrica INAR(1)

A distribuição aqui foi parametrizada pela média e $k \in 0, 1, 2, 3, ...$ As propriedades do processo segue abaixo:

1. $E(Y_t) = \frac{\mu}{1-\alpha}$
2. $Var(Y_t) = \frac{\alpha\mu + \frac{\mu}{(\mu+1)^2}}{(1-\alpha)^2}$
3. $E(Y_t|Y_{t-1}) = \alpha Y_{t-1} + \mu$
4. $Var(Y_t|Y_{t-1}) = Y_{t-1}\alpha(1-\alpha) + \frac{\mu}{(\mu+1)^2}$
5. $Cov(Y_t, Y_{t+j}) = \alpha^j \left(\frac{\alpha\mu + \frac{\mu}{(\mu+1)^2}}{(1-\alpha^2}\right)$, para $j \in N$
6. $\rho_Y(h) = Corr(Y_t, Y_{t+h}) = \alpha^h$, para $h \in N$

2.1.4.2 Estimadores de Mínimos Quadrados Condicionais

Seja $Y_1, Y_2, Y_3, ..., Y_n$ uma amostra do processo Geométrica parametrizada pela média INAR(1) dado na equação (2.3). Estamos interessados em estimar o vetor de parâmetros $\theta = (\alpha, \mu)$. Sabemos que:

$$E(Y_t|Y_{t-1}) = \alpha Y_{t-1} + \mu = g(\theta, Y_{t-1})$$

Considere a função,

$$Q_n(\theta) = \sum_{t=2}^{n} [Y_t - g(\theta, Y_{t-1})]^2$$

Os estimadores de MQC de α e μ são os valores de α e μ que minimizam $Q_n(\theta)$. Depois de derivar $Q_n(\theta)$ em relação a α e μ e sendo igualhada as duas derivadas a zero, obtemos que os estimadores de Minímos quadrados condicionais de α e μ são dados por:

$$\widehat{\alpha}_{MQC} = \frac{\sum_{t=2}^{n} Y_t Y_{t-1} - \frac{1}{n-1}\sum_{t=2}^{n} Y_t \sum_{t=2}^{n} Y_{t-1}}{\sum_{t=2}^{n} Y_{t-1}^{n} Y_{t-1}^2 - \frac{1}{n-1}\left(\sum_{t=2}^{n} Y_{t-1}\right)^2} \tag{2.8}$$

$$\widehat{\mu}_{MQC} = \frac{1}{n-1}\left(\sum_{t=2}^{n} Y_t - \widehat{\alpha}_{MQC}\sum_{t=2}^{n} Y_{t-1}\right) \tag{2.9}$$

2.1.4.3 Estimadores de máxima verossimilhança condicional

Para o estimador de máxima verossimilhança condicional, temos que $Y_t = \alpha \circ Y_{t-1} + \epsilon_t$, em que $\epsilon_t \sim$ Geométrica (k_t, μ_t)

O operador *thinning binomial*: $\alpha \circ Y = \sum_{K=0}^{y} B_K, \quad B_K \sim$ Bernoulli (α).

A função de log-verossimilhança é dada por:

$$l(\alpha, \mu) = \sum_{t=2}^{T} ln \;\; P(Y_t = y_t | Y_{t-1} = y_{t-1})$$

A probabilidade de Transição: $P(y_t|y_{t-1})$ é:

$$\begin{aligned} P(Y_t = y_t | Y_{t-1} = y_{t-1}) &= P(\alpha \circ Y_{t-1} + \epsilon_t | Y_{t-1} = y_{t-1}) \\ &= P(\alpha \circ Y_{t-1} + \epsilon_t = y_t) \end{aligned}$$

Em que $\alpha \circ Y_{t-1} \sim$ Binomial (y_{t-1}, α) e $\epsilon_t \sim$ Geométrica (k, μ), com $X \perp Y$, então,

$$\begin{aligned} P(X + Y) &= \sum_{j=0}^{k} P(X = j, Y = k - j) \\ &= \sum_{j=0}^{k} P(X = j) P(Y = k - j) \end{aligned}$$

Note que, $0 \leq j \leq n$ e $0 \leq j \leq k$. Logo, $0 \leq j \leq min(n.k)$

Logo,

$$\begin{aligned} P(X + Y = k) &= \sum_{j=0}^{min(n,k)} = P(x = j) P(Y = k - j) \\ &= \sum_{j=0}^{min(n,k)} \binom{n}{j} \alpha^j (1 - \alpha)^{n-j} \frac{\mu^{k-j}}{(\mu + 1)^{k-j+1}} \end{aligned}$$

Portanto, a função de log-verossimilhança de um processo geométrica INAR (1) parametrizado por μ é dado por:

$$P(Y_t = y_t | Y_{t-1} = y_{t-1}) = \sum_{j=0}^{min(y_{t-1}, y_t)} \binom{y_{t-1}}{j} \alpha^j (1 - \alpha)^{y_{t-1}-j} \frac{\mu^{y_t - j}}{(\mu + 1)^{y_t - j + 1}}$$

$$l(\alpha, \mu) = \sum_{t=2}^{T} ln \left(\sum_{j=0}^{Min(y_t; y_{t-1})} \binom{y_{t-1}}{j} \alpha^j (1 - \alpha)^{y_{t-1}-j} \frac{\mu^{y_t - j}}{(\mu + 1)^{y_t - j + 1}} \right) \tag{2.10}$$

2.1.5 Processo Binomial Negativa INAR(1)

Considere $\{Y_t, t \in Z\}$ um processo INAR(1), ou sejam um processo em que Y_t satisfaz a equação (2.3). Quando $\epsilon_t, t \in Z$ é um conjunto de variáveis aleatórias independentes com distribuição Binomial Negativa parametrizada pela média, então $\{Y_t, t \in Z\}$ é chamado de processo Binomial Negativa INAR(1). Temos os seguintes resultados, considerando uma distribuição parametrizada pela média:

$$p = \frac{\mu}{\sigma^2}$$

$$P(x = k) = \binom{k + \frac{\mu^2}{\sigma^2 - \mu} - 1}{k} \left(1 - \frac{\mu}{\sigma^2}\right)^k \left(\frac{\mu}{\sigma^2}\right)$$

$\mu^2(\sigma^2 - \mu)$

$$r = \frac{\mu^2}{\sigma^2 - \mu}$$

$$E(X) = \mu$$

$$Var(X) = \sigma^2$$

2.1.5.1 Propriedades do Processo Binomial Negativa INAR(1)

A distribuição foi parametrizada pela média, e as propriedades do processo segue abaixo:

1. $E(Y_t) = \frac{\mu}{1-\alpha}$
2. $Var(Y_t) = \frac{\alpha\mu+\sigma^2}{1-\alpha^2}$
3. $E(Y_t|Y_{t-1}) = \alpha Y_{t-1} + \mu$
4. $Var(Y_t|Y_{t-1}) = y_{t-1}\alpha(1-\alpha) + \sigma^2$
5. $Cov(Y_t, Y_{t+j}) = \alpha^j \left(\frac{\alpha\mu+\sigma^2}{1-\alpha^2}\right), \quad$ para $j \in N$
6. $\rho_Y(h) = Corr(Y_t, Y_{t+h}) = \alpha^h, \quad$ para $h \in N$

2.1.5.2 Estimadores de Minimos Quadrados Condicionais

Seja $Y_1, Y_2, Y_3, Y_4, \ldots, Y_n$ uma amostra do processo Binomial Negativa parametrizada pela média INAR(1) dado na equação (2.3). Estamos interessados em estimar o vetor de parâmetros $\theta = (\alpha, \mu)$. Sabemos que:

$$E(Y_t|Y_{t-1}) = \alpha Y_{t-1} + \mu = g(\theta, Y_{t-1})$$

Considere a função,

$$Q_n(\theta) = \sum_{t=2}^{n} [Y_t - g(\theta, Y_{t-1})]^2$$

Os estimadores de MQC de α e μ são os valores de α e μ que minimizam $Q_n(\theta)$. Depois de derivar $Q_n(\theta)$ em relação a α e μ e sendo igualhada as duas derivadas a zero, obtemos que os estimadores de Minímos Quadrados Condicionais de α e μ são dados por:

$$\widehat{\alpha}_{MQC} = \frac{\sum_{t=2}^{n} Y_t Y_{t-1} - \frac{1}{n-1}\sum_{t=2}^{n} Y_t \sum_{t=2}^{n} Y_{t-1}}{\sum_{t=2}^{n} Y_{t-1}^{n} Y_{t-1}^{2} - \frac{1}{n-1}\left(\sum_{t=2}^{n} Y_{t-1}\right)^2} \tag{2.11}$$

$$\widehat{\mu}_{MQC} = \frac{1}{n-1}\left(\sum_{t=2}^{n} Y_t - \widehat{\alpha}_{MQC}\sum_{t=2}^{n} Y_{t-1}\right) \tag{2.12}$$

2.1.5.3 Estimadores de máxima verossimilhança condicional

Para o estimador de máxima verossimilhança condicional, temos que $Y_t = \alpha \circ Y_{t-1} + \epsilon_t$, em que $\epsilon_t \sim$ Bin. negativa (r_t, P_t)

O operador *thinning binomial*: $\alpha \circ Y = \sum_{K=0}^{y} B_K, \quad B_K \sim$ Bernoulli (α).

A função de log-verosssimilhança é dada por:

$$l(\alpha, \mu, \sigma^2) = \sum_{t=2}^{T} ln \;\; P(Y_t = y_t | Y_{t-1} = y_{t-1})$$

A probabilidade de Transição: $P(y_t|y_{t-1})$ é:

$$\begin{aligned} P(Y_t = y_t|Y_{t-1} = y_{t-1}) &= P(\alpha \circ Y_{t-1} + \epsilon_t | Y_{t-1} = y_{t-1}) \\ &= P(\alpha \circ Y_{t-1} + \epsilon_t = y_t) \end{aligned}$$

Em que $\alpha \circ Y_{t-1} \sim$ Binomial (y_{t-1}, α) e $\epsilon_t \sim$ Bin. Negativa (r, p), com $X \perp Y$, então,

$$\begin{aligned} P(X+Y) &= \sum_{j=0}^{k} P(X = j, Y = k - j) \\ &= \sum_{j=0}^{k} P(X = j) P(Y = k - j) \end{aligned}$$

Note que, $0 \leq j \leq n$ e $0 \leq j \leq k$. Logo, $0 \leq j \leq min(n.k)$

Logo,

$$P(X+Y=k) = \sum_{j=0}^{min(n,k)} = P(x=j)P(Y=k-j)$$

$$= \sum_{j=0}^{min(n,k)} \binom{n}{j}\alpha^j(1-\alpha)^{n-j}\binom{k-j+\frac{\mu^2}{\sigma^2-\mu}-1}{k-j}\left(1-\frac{\mu}{\sigma^2}\right)^{k-j}\left(\frac{\mu}{\sigma^2}\right)^{\frac{\mu^2}{(\sigma^2-\mu)}}$$

Portanto, a função de log-verossimilhança de um processo INAR (1) com distribuição binomial negativa, parametrizada por μ e σ^2 é dada por:

$$P(Y_t=y_t|Y_{t-1}=y_{t-1}) = \sum_{j=0}^{min(y_{t-1},y_t)} \binom{y_{t-1}}{j}\alpha^j(1-\alpha)^{y_{t-1}-j}\binom{y_t-j+\frac{\mu^2}{\sigma^2-\mu}-1}{y_t-j}\left(1-\frac{\mu}{\sigma^2}\right)^{y_t-j}\left(\frac{\mu}{\sigma^2}\right)^{\frac{\mu^2}{(\sigma^2-\mu)}}$$

$$l(\alpha,\mu,\sigma^2) = \sum_{t=2}^{T}\left(\sum_{j=0}^{Min(y_t,y_{t-1})} \binom{y_{t-1}}{j}\alpha^j(1-\alpha)^{y_{t-1}-j}\binom{y_t-j+\frac{\mu^2}{\sigma^2-\mu}-1}{y_t-j}\left(1-\frac{\mu}{\sigma^2}\right)^{y_t-j}\left(\frac{\mu}{\sigma^2}\right)^{\frac{\mu^2}{(\sigma^2-\mu)}}\right) \tag{2.13}$$

2.1.6 Previsão do Processo INAR(1)

Considere a série $Y_1, \cdots, Y_t$ uma amostra do processo INAR (1). Tomando como verdadeiro a premissa de que conhecemos a série até o tempo t, estamos interessados em fazer a previsão de Y_{t+1}.

O valor de k que minimiza o Erro Quadrático Médio Condicional (EQMC), dado por,

$$E\left[(Y_{t+1}-k)^2|Y_t\right] \tag{2.14}$$

é a esperança de Y_{t+1} dado Y_t, ou seja, $k = E[Y_{t+1}|Y_t]$, e usaremos para fazer a previsão de Y_{t+1}.

O valor de a que minimiza o Erro Absoluto Médio Condicional (EAMC), dado por,

$$E = [|Y_{t+1}-a||Y_t] \tag{2.15}$$

corresponde à mediana da distribuição condicional de Y_{t+1} dado Y_t. Então, usaremos a mediana de Y_{t+1} dado Y_t como previsão do valor de Y_{t+1}.

2.1.7 Média da distribuição condicional um passo a frente

Como foi observado na seção anterior, a esperança condicional de Y_{t+1} dado Y_t minimiza o EQMC. Assim, podemos usar esta esperança como previsão de Y_{t+k} dado que conhecemos a série até Y_t. Em sua tese de Doutorado, Freeland (1998), demonstrou que:

$$E(Y_{t+k}|Y_t) = \alpha^k Y_t + \mu_\epsilon \frac{1-\alpha^k}{1-\alpha}$$

Substituindo k por 1, temos que:

$$E\left(Y_{t+1}|Y_t\right) = \alpha Y_t + \mu_\epsilon$$

Usaremos esta expressão para realizar a previsão de Y_{t+1} dado que conhecemos Y_t. Deve-se notar que esta expressão não é necessariamente u número inteiro. Logo, devemos transformar esta esperança para ser um inteiro não negativo, nesse caso o valor mais próximo do número inteiro, e será denotado por $\hat{Y}_{t+1}$.

2.1.8 Mediana da distribuição condicional um passo a frente

Como a mediana da distribuição condicional de Y_{t+1} dado Y_t minimiza o EAMC, podemos usar esta mediana como previsão para Y_{t+1} já que a série até o tempo t é conhecida. Seja $f_1(m|j)$ a função de probabilidade condicional de Y_{t+1} dado $Y_t = k$ é definida como o menor inteiro não-negativo l tal que $\sum_{i=0}^{l} f_1\left(i|k\right) \geq 0.5$. Observe que a mediana da distribuição condicional é pela própria definição um inteiro não-negativo. Este valor será denotado por $\tilde{Y}_{t+1}$.

2.1.9 Medidas de avaliação

O estudo será dividido em duas partes, na primeira em que usaremos para estimar os parâmetros e verificador o desempenho dos métodos de estimação usaremos o viés e o Erro Quadrático Médio (EQM). Considere um conjunto de K simulações de Monte Carlo de amostras de tamanho n do processo INAR(1) e seja $\hat{\theta}^{(i)}$ a estimativa de θ na i-ésima repetição, então o EQM simulado e o viés simulado de $\hat{\theta}$ são dados respectivamente por:

$$EQM(\hat{\theta}) = \frac{1}{K}\sum_{i=1}^{K}\left(\hat{\theta^{(i)}} - \theta\right)^2 \tag{2.16}$$

$$vies(\hat{\theta}) = \theta - \hat{\theta}^*, \tag{2.17}$$

em que $\hat{\theta}^*$ é a estimativa média de θ, ou seja, $\hat{\theta}^* = \frac{1}{K}\sum_{i=1}^{K}\hat{\theta}^{(i)}$.

Na segunda parte do estudo, onde faremos a previsão utilizando os parâmetros encontrados, usaremos o MAE (Erro Médio Absoluto) e a Raiz do Erro Quadrático Médio (RMSE), para avaliar o desempenho dos preditores. Estas medidas podem ser encontradas já sendo utilizadas em trabalhos, como em Mahmoudi, Rostami e Roozegar (2018) para a previsão 1 passo à frente. Seja $Y_1, Y_2, \cdots, Y_{n+m}$ uma amostra aleatória de tamanho $n+m$. A amostra será dividida em duas partes. A primeira parte será formada pelas primeiras n observações, e será utilizada para realizar a estimação dos parâmetros α e λ no processo Poisson, e nos outros casos para realizar a estimação de α e μ, já que estamos usando a distribuição binomial negativa e geométrica parametrizada pela média. A segunda parte

será formada pelas m observações restantes que são usadas para calcular o MAE e o RMSE, que são respectivamente:

$$MAE = \frac{1}{m-1} \sum_{i=n}^{n+m-1} |Y_{i+1} - \hat{Y}_{i+1}| \tag{2.18}$$

$$RMSE = \sqrt{\frac{1}{m-1} \sum_{i=n}^{n+m-1} (Y_{i+1} - \hat{Y}_{i+1})^2} \tag{2.19}$$

onde $\hat{Y}_{i+1}$ representa a previsão de Y_{i+1} dado que conhecemos Y_i.

3 Metodologia

Além da linguagem SQL para realizar a consulta de extração da base de dados, a linguagem R também foi empregada. Todo o processo de manipulação, extração e visualização dos dados foi realizada utilizando o pacote *Tidyverse* e o *Tidymodels*.

O tidyverse é um conjunto de pacotes de software na linguagem de programação R, desenvolvido para facilitar e melhorar a manipulação, análise e visualização de dados. Ele é baseado na filosofia de que a estrutura dos dados deve ser organizada de forma "arrumada" (*tidy*), facilitando o processo de análise (**??**).

Já o tidymodels é um ecossistema de pacotes R projetados para simplificar e padronizar o processo de modelagem de dados. Assim como o tidyverse, o tidymodels segue os princípios de organização e padronização dos dados, tornando mais fácil a criação, avaliação e implantação de modelos estatísticos e de aprendizado de máquina (**??**). Vale lembrar que, o tidymodels faz uso de pacotes já existentes, ele apenas os organiza de uma maneira que seja mais fácil para os usuários trabalharem com eles de forma integrada e seguindo os princípios de organização e padronização de dados do tidyverse.

Todos os códigos podem ser encontrados em minha página no meu github.

3.1 Estudo de Simulação de Monte Carlo

A simulação de Monte Carlo é uma técnica estatística amplamente utilizada para investigar o comportamento de estimadores ou processos estocásticos por meio da repetição de experimentos aleatórios. O método consiste na geração de amostras aleatórias sucessivas, baseadas em distribuições probabilísticas previamente definidas, permitindo a análise empírica da variabilidade dos resultados. Segundo (KROESE et al., 2014), essa abordagem é particularmente útil quando soluções analíticas são impraticáveis ou inexistentes, sendo aplicada em diversas áreas como estatística, física, engenharia e finanças. Tal método fornece uma base robusta para avaliar a precisão, viés e erro quadrático médio de estimadores, especialmente em modelos complexos ou não-lineares (RUBINSTEIN; KROESE, 2016).

Neste trabalho, utilizou-se a simulação de Monte Carlo com o objetivo de estudar o comportamento dos estimadores do modelo INAR(1) sob as três distribuições que o trabalho pretende comparar. Foram geradas milhares de réplicas de séries temporais sintéticas, variando os parâmetros de interesse e avaliando o desempenho dos métodos de estimação adotados. A linguagem `R` foi empregada na construção dos algoritmos de simulação, por ser uma ferramenta poderosa e versátil em estatística computacional R-base. Com ela, foi possível automatizar os procedimentos de geração de dados, estimação e cálculo

das métricas de desempenho, garantindo reprodutibilidade e eficiência no desenvolvimento das análises.

3.1.1 Simulação de Monte Carlo para comparar o desempenho dos estimadores

A simulação tem como objetivo comparar o desempenho dos dois estimadores, Mínimos Quadrados Condicionais (MQC) e Máxima Verossimilhança Condicional (MVC) dos parâmetros α e λ do processo INAR(1), utilizando as três distribuições que vamos comparar. Para isso, foram geradas 5000 amostras do processo INAR(1) para cada distribuição definido na equação (2.3). Para cada amostra gerada calculamos as estimativas de $\widehat{\alpha}_{MQC}$, $\widehat{\alpha}_{MVC}$, $\widehat{\lambda}_{MQC}$ e $\widehat{\lambda}_{MVC}$ com tamanhos amostrais n de 50, 100, 300 e 500 e os seguintes valores dos parâmetros $\alpha = 0.1$ e $\lambda = 5$. Para cada parâmetro, $\widehat{\alpha}$ e $\widehat{\lambda}$ de α e λ, respectivamente, calculamos a estimativa média, o viés e o Erro Quadrático Médio, no caso da distribuição *Binomial Negativa*, utilizamos o parâmetro de variabilidade σ^2.

Tabela 1 – Estimativas dos parâmetros, viés e Erro Quadrático Médio quando $\alpha = 0,1$ e $\lambda = 5$

Estimativa dos Parâmetros												
Amostra	Poisson				Binomial Negativa				Geométrica			
	$\widehat{\alpha}_{MQC}$	$\widehat{\lambda}_{MQC}$	$\widehat{\alpha}_{MVC}$	$\widehat{\lambda}_{MVC}$	$\widehat{\alpha}_{MQC}$	$\widehat{\lambda}_{MQC}$	$\widehat{\alpha}_{MVC}$	$\widehat{\lambda}_{MVC}$	$\widehat{\alpha}_{MQC}$	$\widehat{\lambda}_{MQC}$	$\widehat{\alpha}_{MVC}$	$\widehat{\lambda}_{MVC}$
$n = 50$	0,1002	5,0033	0,1087	4,9560	0,0937	5,0397	0,1080	4,9578	0,0978	5,0047	0,1022	4,9823
$n = 100$	0,0969	5,0185	0,0997	5,0027	0,0984	4,9929	0,1055	4,9540	0,0975	5,0125	0,0998	5,0032
$n = 300$	0,0964	5,0192	0,0972	5,0147	0,0977	5,0116	0,1020	4,9877				
$n = 500$	0,0979	5,0121	0,0984	5,0089	0,0972	5,0163	0,1011	4,9942				
Viés												
$n = 50$	0,0002	0,0033	0,0087	-0,0440	-0,0063	0,0397	0,0080	-0,0422	-0,0022	0,0047	0,0022	-0,0157
$n = 100$	-0,0031	0,0185	-0,0003	0,0027	-0,0016	-0,0071	0,0055	-0,0460	-0,0025	0,0125	-0,0002	0,0032
$n = 300$	-0,0036	0,0192	-0,0028	0,0147	-0,0023	0,0116	0,0020	-0,0123				
$n = 500$	-0,0021	0,0121	-0,0016	0,0089	-0,0028	0,0163	0,0011	-0,0058				
Erro Quadrático Médio												
$n = 50$	0,0107	0,4316	0,0129	0,4971	0,0108	0,9802	0,0036	0,7035	0,0106	0,4475	0,0111	0,4626
$n = 100$	0,0069	0,2593	0,0073	0,2722	0,0073	0,5234	0,0018	0,3535	0,0070	0,2661	0,0070	0,2650
$n = 300$	0,0030	0,1072	0,0030	0,1072	0,0031	0,1917	0,0006	0,1120				
$n = 500$	0,0019	0,0683	0,0019	0,0680	0,0020	0,1217	0,0003	0,0695				

Fonte: Elaboração própria

Na tabela 1 em todas as distribuições, com a combinação de parâmetros que foi definido, o EQM de cada um dos estimadores a medida que n aumenta tende para 0. Isto evidencia que os estimadores $\widehat{\alpha}_{MQC}$, $\widehat{\lambda}_{MQC}$, $\widehat{\alpha}_{MVC}$ e $\widehat{\lambda}_{MVC}$ são consistentes[1].

[1] Um estimador é consistente quando, à medida que o tamanho da amostra aumenta, ele converge em probabilidade para o valor verdadeiro do parâmetro que está sendo estimado. Em outras palavras, $\hat{\theta}_n \xrightarrow{P} \theta$ quando $n \to \infty$.

Tabela 2 – Amostra da base de dados bruta

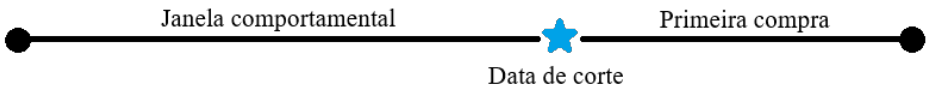
CODCLI	VLATEND	DATA
1033	R$ 144,18	03/05/2013
958	R$ 477,00	06/05/2013
1220	R$ 447,45	06/05/2013
1144	R$ 449,60	06/05/2013
1213	R$ 551,00	01/02/2013

Fonte: Base de dados da distribuidora

3.2 Pré Processamento

Antes de realizar a seleção das variáveis que iriam compor o modelo, foi definida uma data de corte. O dia 1 de janeiro de 2024 foi utilizado para repartir os dados em duas partes. A primeira, chamamos de "janela comportamental", na qual comportou pouco mais de 11 anos do histórico de compra dos clientes, e a segunda como "primeira compra", responsável por armazenar temporariamente, como o próprio nome já diz, a data da primeira compra após a janela comportamental.

Figura 1 – Estrutura dos Dados de Treinamento



Fonte: Elaboração própria

3.2.1 Segmentação RFV

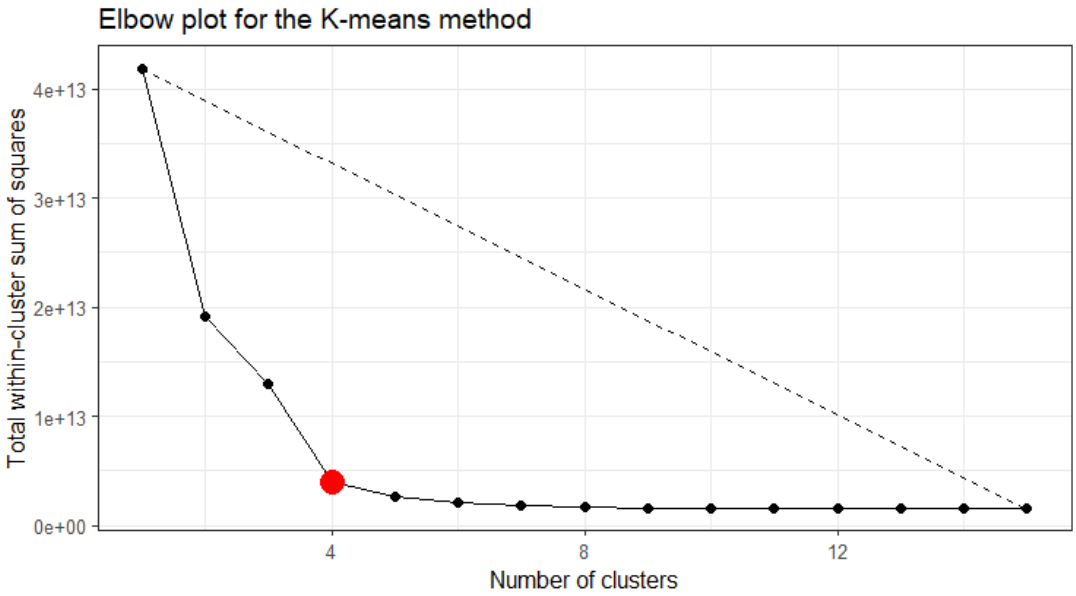
A segmentação RFV é comumente utilizada por profissionais de marketing para classificar clientes com base em seu comportamento de compra. A sigla representa: Recência, Frequência e Valor Monetário. Recência refere-se à última vez que um cliente fez uma compra, isto é, o quão ativo o cliente está. A frequência determina o quanto um cliente pode ser leal. E por fim, o valor monetário representa clientes que contribuem significativamente para a receita de uma empresa (**??**).

Analisando esses três aspectos em conjunto, podemos segmentar os clientes em diferentes grupos: "clientes VIP", "clientes que não podemos perder", "clientes inativos que poderíamos reativar", dentre muitas outras classificações.

Para determinar os *clusters* foi utilizada uma técnica de aprendizagem não supervisionada, *K-Means*. O objetivo é agrupar os dados de tal forma que os pontos dentro de cada cluster sejam semelhantes entre si e diferentes dos pontos em outros *clusters*. O

método assume que a medida de dissimilaridade usada é a distância Euclidiana. Para utilizar, é necessário especificarmos de antemão o valor de K, quantos *clusters* desejamos identificar. Para isso, utilizamos o método do cotovelo, de maneira objetiva, o método do cotovelo baseia-se na ideia de que a medida em que o número de *clusters* aumenta, a variabilidade *intra-cluster* diminui, levando a uma diminuição na soma dos quadrados das distâncias *intra-cluster*.

Figura 2 – Número ótimos de cluster pela Técnica do Cotovelo



Fonte: Elaboração própria

Na primeira etapa, foi aplicada a técnica *K-Means* de maneira individual para cada uma das variáveis mencionadas anteriormente e formando *clusters* para cada uma delas. Em outras palavras, foram formados quatro grupos de clientes que apresentaram recência semelhante, quatro grupos de clientes com frequência semelhantes e, por fim, quatro grupos de clientes com valor monetário gasto semelhante. Na segunda etapa, foram atribuídos pontuações de zero a três para cada um dos grupos, baseando-se em seus valores originais. Por exemplo, grupos de clientes que gastaram pouco dinheiro na empresa receberam a pontuação zero, enquanto grupos de clientes que mais gastaram dinheiro receberam a pontuação três. Por fim, na terceira etapa, somamos essas pontuações e, através dos quartis, atribuímos as seguintes classificações: “Baixo Valor”, “Médio Valor” e “Alto Valor”.

3.2.2 Estrutura final da base de dados

Após todo o processo de engenharia de características, o banco de dados final ficou composto por 17 variáveis e 4521 observações (clientes).

Tabela 3 – Identificação das Variáveis

Variável	**Identificação**
dcadastro	Total de dias desde a primeira compra
d1, d2 e d3	Diferença entre os dias de compra das últimas 4 compras
dmedia	Média da diferença entre os dias de compra
ddesvio	Desvio Padrão da diferença entre os dias de compra
dmedian	Mediana da diferença entre os dias de compra
r	Diferença de dias entre a última compra e a data corte
f	Contagem do número de pedidos realizados
v	Valor total gasto
clr, clf e clv	*clusters* da recência, frequência e valor
RFV	Soma das pontuações dos *clusters*
clusters	Clientes de "baixo", "médio", "alto valor"
Y	Primeira compra após o corte menos a última compra antes do corte

Fonte: Elaboração própria

3.3 Validação

Para avaliar o poder preditivo de um modelo, isto é, a capacidade do modelo estimar valores ou classificar de maneira correta novas observações, é necessário dividirmos nosso banco de dados em um conjunto de treinamento e um conjunto de teste. Normalmente, a proporção escolhida para repartição é selecionada de forma empírica, isto é, a depender do usuário em questão. Para o projeto em questão, foi selecionada a proporção de 70/30 para o treino e teste, respectivamente. Essa divisão é extremamente necessária, deve ser feita de maneira aleatória, para evitar sobreajuste, isto é, quando o modelo se ajusta muito bem aos nossos dados de treino e não consegue prever muito bem os resultados de novas observações.

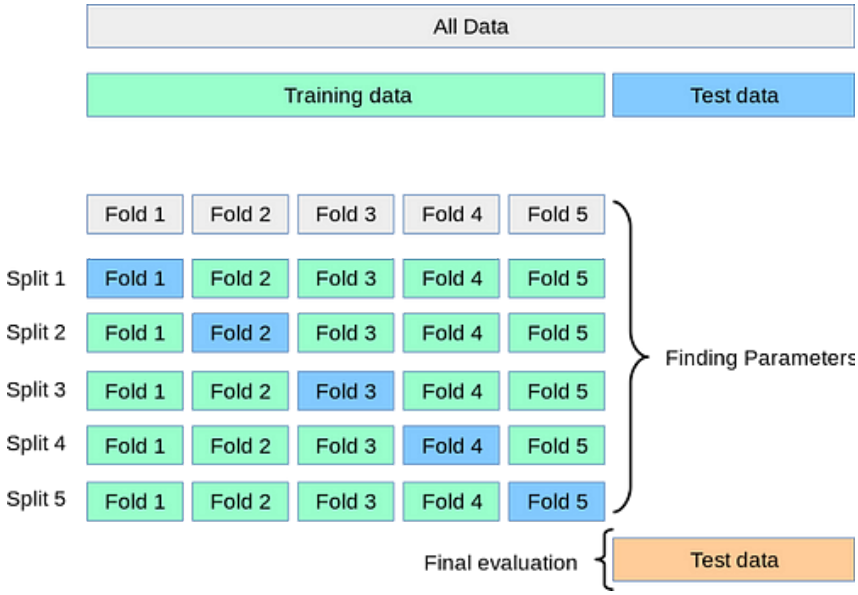
3.3.1 Validação Cruzada

A técnica de validação utilizada no projeto chama-se *K-fold*. De acordo com **??**), nessa abordagem o conjunto de dados de treinamento é repartido em K subgrupos de maneira aleatória e com aproximadamente o mesmo tamanho. Em seguida, é treinado um modelo utilizando K-1 subgrupos e utilizando o grupo removido como o banco de teste para avaliar o desempenho. No projeto em questão, foi utilizado K = 10, ou seja, o banco de treinamento foi repartido de maneira aleatória em 10 subgrupos. Na figura 3 ilustra esse processo de divisão dos dados e um processo de K-fold, com K = 5.

3.4 Avaliação

A medida utilizada para avaliar os modelos em questão foi o erro quadrado médio (EQM), como o conjunto de teste não é utilizado para estimar os parâmetros do modelo, o

Figura 3 – Ilustração de um processo de Validação 5-fold Validation



Fonte: Arquivo de Internet

EQM torna-se consistente pela lei forte de Kolmogorov, ou como normalmente é conhecida, lei dos grandes números.

A fórmula do Erro Quadrático Médio (EQM) é dada por:

$$EQM = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.1}$$

onde:

- n é o número de observações.
- y_i representa os valores reais observados.
- $\hat{y}_i$ são os valores previstos pelo modelo.
- $(y_i - \hat{y}_i)^2$ é o quadrado da diferença entre o valor real e o valor previsto para a i-ésima observação.

A lei afirma que, sob certas condições, a média aritmética de uma sequência de variáveis aleatórias independentes e identicamente distribuídas (i.i.d.) converge quase certamente para a esperança matemática (média esperada) dessas variáveis à medida que o número de observações aumenta para o infinito **??**). Em outras palavras a média dos valores estimados converge para a média real a medida que a amostra aumenta. A fórmula da Lei Forte dos Grandes Números é dada por:

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{qc.} \mu \quad \text{quando} \quad n \to \infty \tag{3.2}$$

onde:

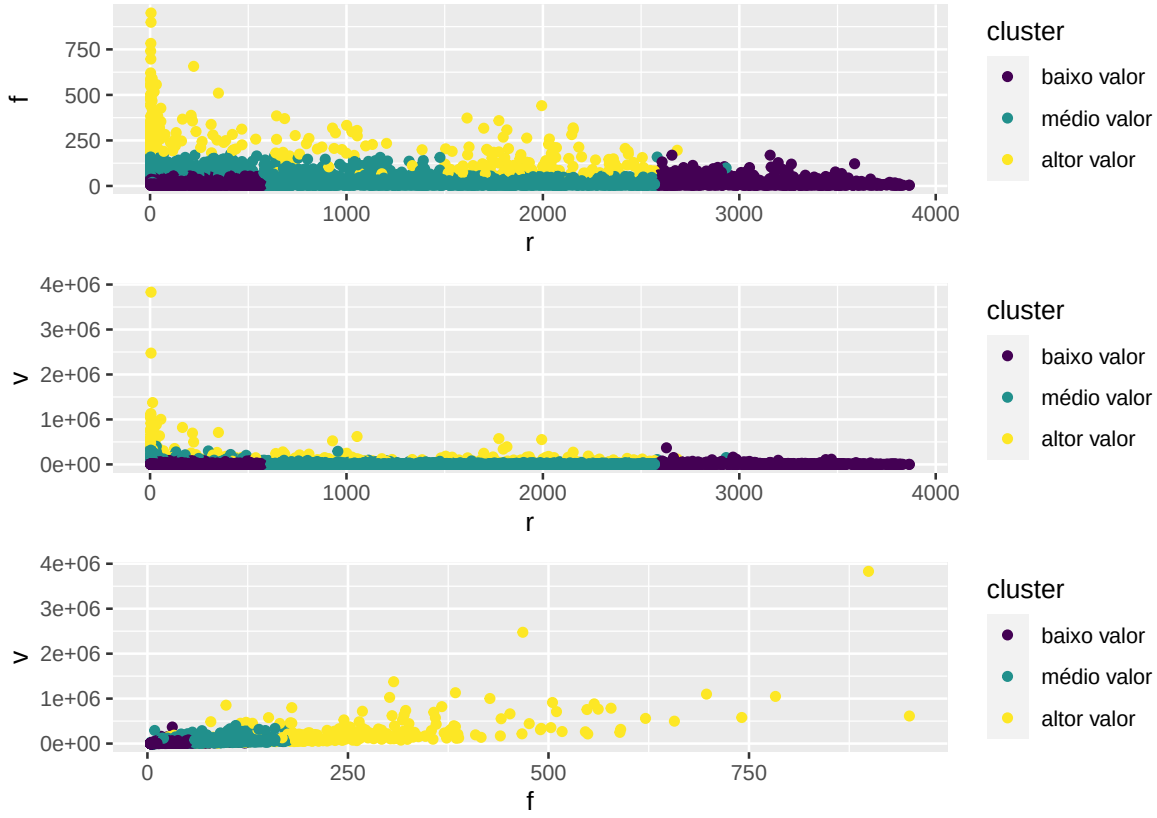
- $X_1, X_2, \ldots$ são variáveis aleatórias independentes e identicamente distribuídas (i.i.d.).
- μ é a esperança matemática das variáveis aleatórias X_i.
- $\xrightarrow{qc.}$ denota convergência quase certa.

4 Resultados

4.1 Análise Exploratória

Os gráficos de dispersão apresentados na Figura 4 exibem os grupos formados a partir da segmentação de clientes utilizando a técnica de agrupamento *K-means*. No grupo de alto valor, é notável um alto valor de gastos e frequência de compras, além de uma baixa e média recência. No grupo de baixo valor, é possível notar de forma contrária o mesmo comportamento.

Figura 4 – Segmentação de clientes - Banco Completo



Fonte: Elaboração Própria

Na Tabela 4, são apresentadas as estatísticas descritivas das variáveis dos grupos definidos. Entre os três grupos identificados, o de "baixo valor" foi o grupo com o maior número de clientes, contando com um total de 2.111. O grupo de "médio valor" ficou em segundo lugar, com 2.072 clientes, enquanto o grupo de "alto valor" ficou com 338 clientes.

Como já era esperado, o grupo de alto valor gerou uma receita média de 212 mil

reais, com uma média de 224 pedidos. Isso representa aproximadamente 9 vezes mais em receita e 6,5 vezes mais em pedidos quando comparado com o grupo de valor médio, que registrou uma média de 23 mil reais em receita e 35 pedidos. É importante ressaltar que a recência média de compras dos três grupos foi mais alta do que o esperado, isso devido à alta taxa de clientes inativos na base de dados completa.

Tabela 4 – Análise Descritivas dos Grupos - Banco Completo

	Grupos (N = 4521)		
	Baixo Valor (N = 2.111)	**Médio Valor (N = 2.072)**	**Alto Valor (N = 308)**
Valor Gasto Médio	11.817,71 R$	23.056,42 R$	212.534,69 R$
Frequência Média	17	35	224
Recência Média	2.111 dias	2.072 dias	338 dias

Fonte: Elaboração Própria

Na tabela 5 temos a distribuição dos clientes que voltaram a realizar compras após a data de corte definida. Nota-se que houve uma diminuição significativa no número de clientes ativos na empresa, 77,3% dos clientes não voltaram a realizar compras. A causa dessa grande inatividade pode se dar por inúmeros fatores, desde motivações externas, como a perda de interesse nos produtos fornecidos ou aumento de preços, até mesmo por uma falha gerencial por parte da empresa que tenha deixado de oferecer seus serviços.

Dos grupos de clientes formados, no grupo de "alto valor" ocorreu uma diminuição de 57,5%, em que caiu de 308 clientes para 131 clientes. Já para o grupo de "médio valor" a diminuição foi de 86,4%, caindo de 2.072 clientes para 283 clientes. Por fim, o grupo em que houve a maior diminuição foram os clientes de "baixo valor", cerca de 71%, que caíram de 2.111 clientes para 612 clientes.

Tabela 5 – Distribuição dos clientes que voltaram a comprar

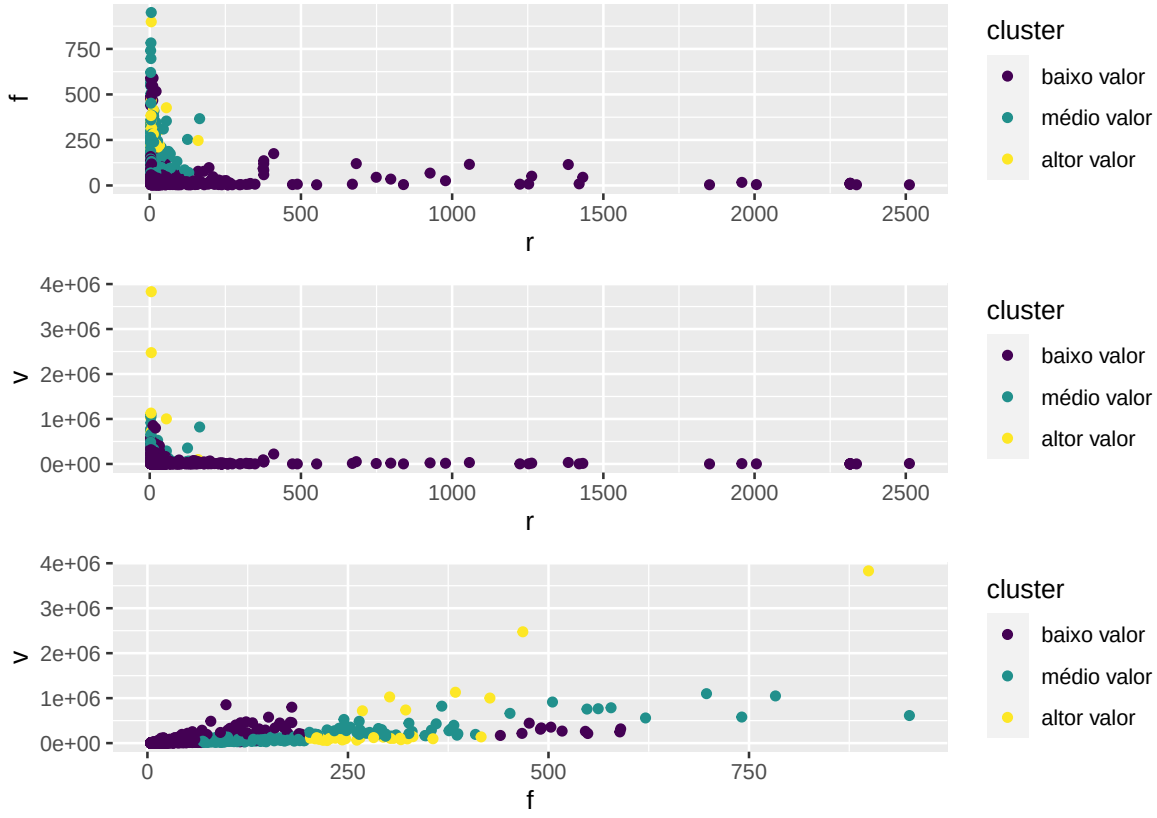
	Sim		**Não**	
Grupos	**f**	**fr (%)**	**f**	**fr (%)**
Baixo Valor	612	13,5%	1499	33,2%
Médio Valor	283	6,3%	1789	39,6%
Alto Valor	131	2,9%	207	4,5%
Total	**1.026**	**22,7%**	**3.495**	**77.3**

Um fato importante a se destacar é que esses clientes que não retornaram a realizar compras foram removidos do banco de dados final. Por se tratarem de observações que necessitavam um tratamento especial, não fazia sentido imputar valores na nossa variável resposta e, por conta disso, será discutido ao final do trabalho.

Na Figura 5 temos o gráfico de dispersão par a par das variáveis usadas na segmentação. Nota-se que a estrutura principal se mantém, entretanto, agora com uma

menor densidade de observações. Nos dois primeiros gráficos, é possível visualizar que muitos dos clientes com recência muito alta foram removidos.

Figura 5 – Segmentação de clientes - Variável resposta faltante omitida



Fonte: Elaboração Própria

A heterogeneidade dos dados entre os grupos se manteve, dessa vez as diferenças entre eles ficaram mais evidentes na tabela 6. O grupo de alto valor, com 343.247,92 reais gastos, cerca de 23 vezes mais que o valor médio gasto do grupo de baixo valor e cerca de 4,5 vezes mais que o valor médio gasto do grupo de médio valor. O mesmo padrão reflete na frequência média de pedidos. A principal mudança está na variável recência, esta agora apresenta valores mais baixos. O grupo de alto valor os clientes apresentaram uma recência média de 30 dias, enquanto os demais grupos a recência gira em torno de 72 dias.

Tabela 6 – Análise Descritivas dos Grupos em que observações com a variável resposta faltante foram omitidas.

	Grupos (N = 1026)		
	Baixo Valor (N = 612)	**Médio Valor (N = 283)**	**Alto Valor (N = 131)**
Valor Gasto Médio	14.907,63 R$	77.562,09 R$	343.247,92 R$
Frequência Média	19	89	312
Recência Média	69	75	30

4.2 Resultados dos Modelos

Um primeiro ajuste foi realizado utilizando-se de todo o conjunto de variáveis disponíveis que haviam sido extraídas. Foi observado que os resultados de treinamento estavam excelentes, porém quando o modelo era avaliado no conjunto de teste, os resultados não eram tão satisfatórios, ocasionados por um provável sobreajuste do modelo.

Isso levou a um processo cansativo para selecionar as variáveis ideais que deveriam compor o modelo. Para isso, foi realizada a extração de 15 subconjuntos de dados. Isto é, foram criados subconjuntos de dados, com a mesma quantidade de observações, mas com o conjunto de variáveis diferentes. Essa abordagem nos permitiu explorar uma variedade de combinações diferentes de variáveis que compusessem a base para reduzir o risco de um ajuste excessivo por parte do modelo. No fim de uma extensa sessão de treinos e testes, o banco de dados que se adequou à proposta era composto pelas variáveis apresentadas na tabela 7:

Tabela 7 – Identificação das Variáveis do Banco Final

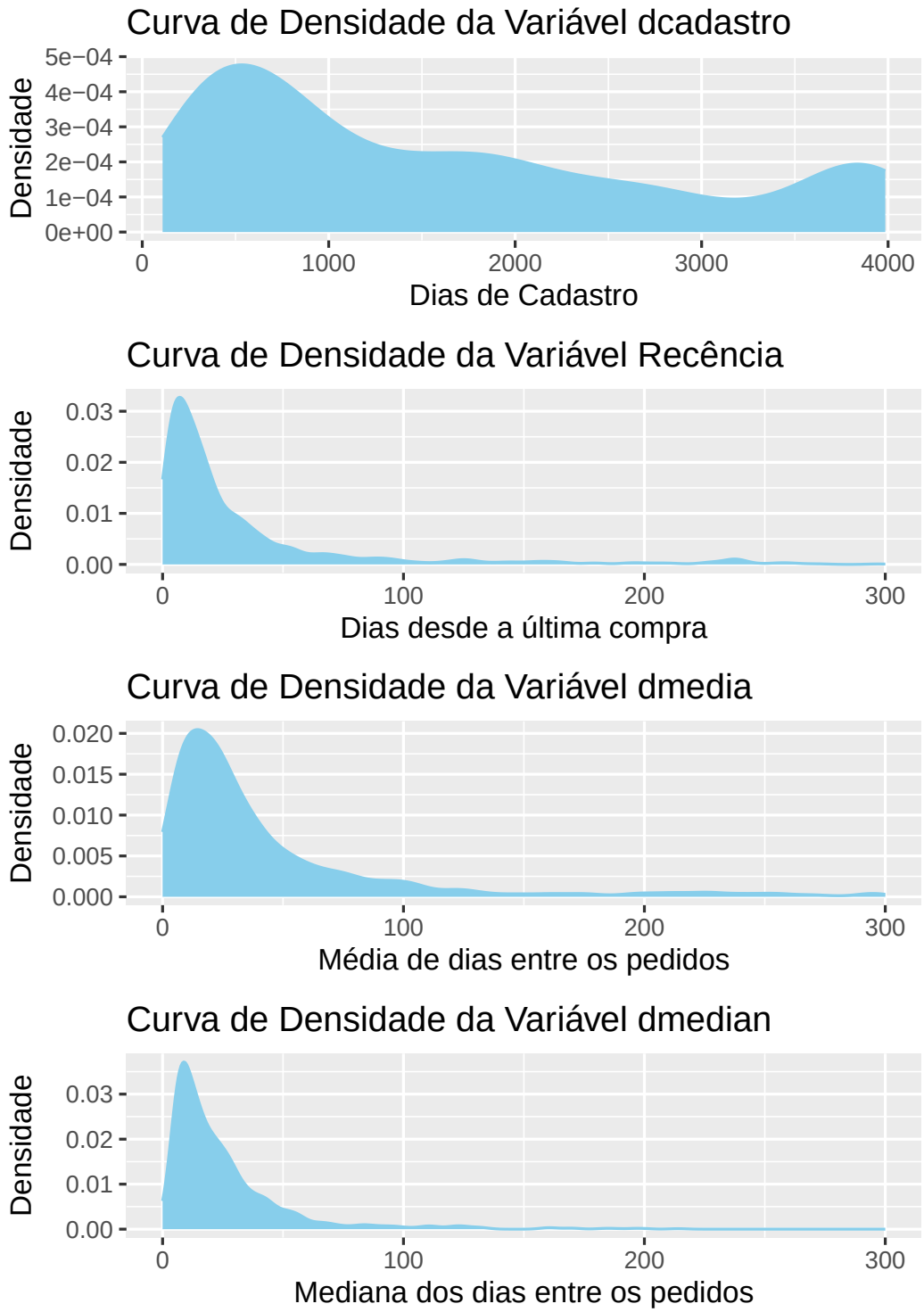
Variável	**Identificação**
dcadastro	Total de dias desde a primeira compra
r	Dias desde a última compra
dmedia	Média da diferença entre os dias de compra
dmedian	Mediana da diferença entre os dias de compra
clusters	Clientes de “baixo”, “médio”, “alto valor”
Y	Dias entre a primeira compra após o corte e a última compra antes do corte

Fonte: Elaboração própria

A figura 6 é apresentado as curvas de densidade das variáveis do banco. A variável *dcadastro* em comparação com as demais variáveis explicativas do nosso banco de dados, é a mais balanceada. Esse formato da curva de densidade é decorrente de uma fidelidade por parte de alguns clientes antigos que ainda efetuam compras na empresa. A densidade da variável *dcadastro* ainda nos mostra que a grande concentração de clientes são aqueles que efetuaram cadastro por volta de 1 ano e meio na empresa. A respeito da densidade das demais variáveis, a assimetria à direita já era esperada, dado que quanto menor fosse o valor dessas variáveis, mais ativos seriam os clientes. Além disso, investigando mais a fundo,

foi possível constatar que alguns clientes antigos passaram longos períodos sem compras e voltaram à atividade e outros realmente de fato inativaram, deixando de comprar.

Figura 6 – Curva de densidade das variáveis do modelo



Fonte: Elaboração própria.

A figura 7 apresenta os melhores resultados dos modelos treinados, medidos pela raiz do erro quadrático médio (RMSE). O modelo *boost tree* (XGB) se destacou apresentando o menor RMSE entre os modelos testados. Em segundo lugar, temos o modelo *mlp* (Perceptron Multicamadas), que também mostrou um bom desempenho, embora ligeiramente inferior ao XGB.

O modelo *bag mlp*, que utiliza uma abordagem de combinar de múltiplos perceptrons multicamadas, ficou em terceiro lugar e embora seu desempenho não tenha sido tão bom quanto os dois primeiros, ele ainda apresentou uma precisão razoável. O modelo *rand forest* (Floresta Aleatória) apareceu na quarta posição. Apesar de as florestas aleatórias combinarem várias árvores de decisão para melhorar a precisão, neste caso, seu RMSE foi maior do que os modelos anteriores, indicando uma menor eficiência.

Por fim, o modelo *svm rbf* (SVM com Kernel RBF) apresentou o pior desempenho, com o RMSE mais alto entre todos os modelos avaliados. Isso sugere que, para este conjunto de dados específico, o SVM não foi capaz de se ajustar tão bem quanto aos demais.

Figura 7 – Rankings dos melhores modelos após a validação cruzada

Fonte: Elaboração própria.

Na tabela 8 podemos observar as métricas de desempenho após avaliado no conjunto teste. No treinamento, o MLP teve um bom desempenho, figurando como o segundo melhor modelo, com um RMSE baixo. No conjunto de teste, o MLP manteve seu excelente desempenho, alcançando o menor EQM dos modelos avaliados e um pseudo-R^2 de 0.985. Esta consistência sugere que o MLP não só foi bem ajustado, mas também generaliza bem para novos dados.

A Floresta Aleatória ficou em quarto lugar no ranking após ser avaliado no conjunto de treinamento, com um desempenho um pouco inferior aos melhores modelos, mas ainda competitivo. No conjunto de teste, o modelo apresentou uma EQM de 20.7 e um pseudo-R^2 de 0.984, mostrando uma leve melhoria em relação ao que se poderia esperar dado seu ranking de treinamento. Este modelo também mostrou boa capacidade de generalização.

O modelo XGB que havia se destacado no ranking de treinamento, não se manteve consistente ao generalizar para novas observações. Entretanto, a diferença entre os 4 primeiros modelos não foi tão discrepante quanto ao SVM com kernel RBF. O mesmo apresentou o pior desempenho nos treinos, o que foi refletido também no conjunto de teste. Com um EQM de 52.6 e um pseudo-R^2 de 0.922, este modelo mostrou-se menos eficaz em se ajustar aos dados.

Tabela 8 – Desempenho dos modelos no conjunto teste

Modelo	**Erro Quadrado Médio**	**pseudo-R^2**
MLP	20	0.985
Floresta Aleatória	20.7	0.984
Bagging MLP	21.6	0.977
XGB	21.7	0.973
Support Vector Machine	52.6	0.922

Fonte: Elaboração própria

5 Considerações finais

O presente trabalho teve como finalidade avaliar modelos de aprendizado de máquina na tentativa de prever o próximo dia de compra de um cliente. Pode-se dizer que, com os resultados obtidos dos modelos avaliados, as técnicas mostraram-se eficazes para prever o próximo dia de compra dos clientes, apresentando resultados satisfatórios para o banco de dados utilizado. Apesar dos desafios iniciais, relacionados ao sobreajuste do modelo aos dados de treinamento, o uso da engenharia de características permitiu alcançar um resultado satisfatório, deixando de aprendizado a real importância dos passos iniciais no treinamento de um modelo. Vale deixar claro que, apesar do modelo *Multilayer Perceptron* ter apresentado o melhor resultado, isso não implica dizer que ele é de fato o melhor modelo, e sim, de que ele foi bom em se ajustar à esses dados em questão. Além disso, exceto pelo modelo SVM, todos os demais modelos apresentaram resultados satisfatórios.

Um ponto importante nesse trabalho que vale se destacar foi a junção da segmentação de clientes RFV com o método de aprendizagem não supervisionada, *K-means*, na qual retorna uma informação importante para o setor comercial. Com base nessa segmentação, as empresas podem ter um controle maior acerca de seus clientes, a possibilidade de personalizar suas estratégias de marketing e comunicação para atender às necessidades específicas de cada grupo de clientes, garantem uma maximização do retorno sobre investimento. A respeito do real objetivo do trabalho, previsão do próximo dia de compra, essa informação por si só possibilitaria o desenvolvimento de ofertas, promoções e mensagens especiais programadas para incentivar a repetição de compras e aumentar o valor do cliente. Além disso, esse dado possibilita um gerenciamento mais eficiente do estoque, evitando excesso ou falta de estoque.

Em suma, os resultados do estudo destacam o potencial da utilização do aprendizado de máquina no setor comercial. Além da previsão do próximo dia de compra, as técnicas podem ser adaptadas para identificar outros comportamentos, como preferências de produtos e ciclos de vida do cliente. Em geral, é importante reconhecer as limitações deste estudo, reiterando que um bom modelo requer boas informações, além-claro, das melhorias no pré-processamento: seleção de variáveis e tratamento das informações.

Por fim, em um certo ponto da metodologia, foi mencionado que algumas observações apresentaram variável resposta faltante, decorrente do fato de alguns clientes não voltarem a realizar compras. Esse tipo de comportamento é comum no campo da análise e sobrevivência. A área lida especificamente com os chamados “dados censurados”, em que a variável resposta, nesse caso uma compra subsequente, pode ser faltante para algumas observações. Futuras pesquisas podem abordar essas limitações para aprimorar esses modelos, com técnicas de análise e sobrevivência ou até mesmo remodelar para um

problema de classificação, buscando identificar clientes que deixarão de comprar em uma janela de tempo pré-definida.

Referências

AL-OSH, M. A.; ALZAID, A. A. First-order integer-valued autoregressive (inar(1)) process. *Journal of Time Series Analysis*, v. 8, n. 3, p. 261–275, 1987.

BARCELOS, B. I. *Teste da raiz unitária Dickey-Fuller no modelo INAR(1)*. Dissertação (Mestrado) — Universidade Federal do Espírito Santo, 2008.

CAMERON, A. C.; TRIVEDI, P. K. *Regression Analysis of Count Data*. [S.l.]: Cambridge University Press, 1998.

DU, J. G.; LI, Y. The integer-valued autoregressive (inar(p)) model. *Journal of Time Series Analysis*, v. 12, n. 2, p. 129–142, 1991.

FOKIANOS, K.; TJøSTHEIM, D. Poisson autoregression. *Journal of the American Statistical Association*, v. 104, n. 488, p. 1430–1439, 2009.

FREELAND, R. K. *Statiscial analysis of discrete time series with application to the analysis of workers compensation claims data*. Tese (Doutorado) — University of British Columbia, 1998.

GOMES, K. S. *Modelagem INAR(p) para previsão de índices de qualidade do ar*. Dissertação (Mestrado) — Universidade Federal do Espírito Santo, 2009.

KROESE, D. P. et al. Why the monte carlo method is so important today. *Wiley Interdisciplinary Reviews: Computational Statistics*, Wiley, v. 6, n. 6, p. 386–392, 2014.

LATOUR, A. Existence and stochastic structure of a non-negative integer-valued autoregressive process. *Journal of Time Series Analysis*, v. 19, n. 4, p. 439–455, 1998.

MAHMOUDI, E.; ROSTAMI, M.; ROOZEGAR, R. A new integer-valued ar(1) process based on power series thinning operator. *Communications in Statistics - Simulation and Computation*, v. 47, n. 10, p. 2895–2906, 2018.

MCKENZIE, E. Some simple models for discrete variate time series. *Water Resources Bulletin*, v. 21, n. 4, p. 645–650, 1985.

MCKENZIE, E. Autoregressive moving-average processes with negative binomial and geometric marginal distributions. *Advances in Applied Probability*, v. 18, n. 3, p. 679–705, 1986.

MCKENZIE, E. Discrete variate time series. In: FINKENSTäDT, B.; ROOTZéN, H. (Ed.). *Seminar on Stochastic Analysis, Random Fields and Applications III*. [S.l.]: Springer, 2003. p. 305–312.

PEDELI, X.; KARLIS, D. A bivariate inar(1) model for count time series. *Statistical Modelling*, v. 11, n. 1, p. 35–53, 2011.

R, Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2025. Disponível em: <https://www.R-project.org/>.

RUBINSTEIN, R. Y.; KROESE, D. P. *Simulation and the Monte Carlo Method*. 3. ed. [S.l.]: John Wiley & Sons, 2016.

SILVA, I. M. M. d. *Contributions to the analisys of discrete-valued time series*. Tese (Doutorado) — Universidade Federal do Rio Grande do Norte, 2005.

STEUTEL, F. W.; HARN, K. van. Discrete analogues of self-decomposability and stability. *The Annals of Probability*, v. 7, n. 5, p. 893–899, 1979.

ZHANG, J.; WANG, M.; ZHU, F. Modeling monthly tuberculosis incidence with inar models. *Statistics in Medicine*, v. 38, n. 11, p. 1953–1966, 2019.