# Machine Learning aplicado para análise de comportamento de consumo e previsão de evasão de cliente

Letícia Junqueira Inglez de Souza[1]*; Daniele Aparecida Cicillini Pimenta[2]

[1] PECEGE, USP Bacharel em Administração de Empresa. Rua Padre João Manoel da Silva, 413 – Nova América; 13417-770 Piracicaba, São Paulo, Brasil

[2] PECEGE, USP. Mestre em Engenharia. Rua Padre João Manoel da Silva, 413 – Nova América; 13417-770 Piracicaba, São Paulo, Brasil

*autora correspondente: leticiainglez@hotmail.com

# “Machine Learning” aplicado para análise de comportamento de consumo e previsão de evasão de cliente

## Resumo

A evolução da era digital, as mudanças nas formas de pessoas e empresas se relacionarem e consumirem, aliadas ao avanço dos volumes de dados (Big Data) obrigaram as empresas a buscarem soluções para compreender e reter seus clientes. A forma pela qual as empresas fazem isso se dá por meio da utilização de modelos de inteligência artificial e aprendizado de máquina, antecipando-se a possíveis evasões. O desenvolvimento, de forma ágil, as comunicações e os produtos orientados para as necessidades dinâmicas dos clientes são mecanismos que tornam as empresas mais competitivas e atrativas e competitivas dentro do mercado. Este estudo propõe uma maneira de identificar o relacionamento com o negócio, a previsão de abandono e a segmentação de comportamento para antecipar ações estratégicas de marketing a fim de minimizar a evasão de clientes.

**Palavras-chave:** Aprendizado de Máquina; “churn”; Algoritmo; Predição; Inteligência Artificial.

## Abstract

With the evolution of the digital age, the change in the way people and companies relate and consume, added to the advance of data volumes (Big Data), forced companies to seek solutions to understand and retain their customers. The way in which companies perform such solutions is through the use of artificial intelligence and “Machine Learning” models, anticipating possible evasions. Agile development, communications and products oriented towards the dynamic needs of customers are mechanisms that make companies more competitive and attractive within the market. This study proposes a way to identify the relationship with the business, predict abandonment and segmentation of behavior to anticipate strategic marketing actions in order to minimize customer evasion.

**Keywords:** “Machine Learning”; “churn”; Algorithm; Prediction; Artificial Intelligence.

## Introdução

Em meio a constantes mudanças, o mundo atualmente passa pela quarta revolução industrial, representada pela integração de diferentes tecnologias, tais como: inteligência artificial, robótica, internet das coisas e computação em nuvem, todas com o objetivo de

promover a digitalização das atividades empresariais, otimizar os processos e aumentar a produtividade. (Schwab, 2018)

O avanço exponencial da tecnologia trouxe também o Big Data, ou seja, grandes e diversos volumes de dados que chegam diariamente com alta velocidade nas empresas, que armazenam essas informações em seus servidores e bancos de dados. O Big Data é considerado o novo petróleo das empresas. No entanto, para que possam ter realmente valor, os dados precisam ser processados, interpretados e transformados em conhecimento, gerando, assim, valor para as empresas. (Magaldi, 2018)

Os grandes volumes de dados aliados às tecnologias de inteligência artificial e de aprendizado de máquina promovem, entre outras soluções, a compreensão do comportamento do consumidor, facilitando a personalização e customização de ofertas e campanhas de marketing e influenciando as estratégias das empresas. (Kotler, 2017)

Dentro desse novo panorama, as organizações têm direcionado seus esforços em estratégias com foco no cliente (“Client-centric”) com o propósito de alinhar seus produtos e serviços às necessidades e aos desejos atuais e futuros de seus clientes mais valiosos para maximizar os lucros no longo prazo. Da mesma forma, essas organizações estão concentradas em criar a melhor experiência com o intuito de fidelizar seus clientes à marca e mantê-los satisfeitos. (Kotler, 2017)

Consequentemente, as empresas têm utilizado os avanços tecnológicos e técnicas de aprendizado de máquina (““Machine Learning””) para o desenvolvimento de soluções e tomadas de decisões baseada em dados, a fim de elevar seus negócios a um nível mais competitivo, pois os consumidores estão cada vez mais atentos e exigentes às ofertas de produtos, preços e atendimentos personalizados.(Fawcett, 2013)

Clientes cada vez mais conscientes, exigentes e voláteis (propícios a trocar de empresa, produto, serviço) tornam a competitividade entre empresas mais intensa quanto à busca de dados, necessidades, preferências, hábitos de consumo e demais informações de seus clientes a fim de obter maior compreensão, segmentação dos perfis e desenvolver estratégias de relacionamentos baseados na confiança para conseguir retê-los, encantá-los e fidelizá-los de tal forma que o relacionamento com a marca perdure. (Kotler, 2021)

É crescente a preocupação das organizações em relação à gestão de relacionamento com os seus consumidores, mais especificamente com o abandono/evasão por parte deles, ou seja, o chamado “churn”, termo que se refere a clientes que consomem produtos e/ou serviços mas que, por alguma razão, deixam de adquirir daquela empresa e passam a consumir da concorrência.

O “churn” também está diretamente associado ao tempo em que o indivíduo permanece como cliente de uma organização, ou seja, relaciona-se com o conceito de

"Consumer Life Time Value" (valor do ciclo de vida do cliente), que se traduz pela mensuração do lucro, presente e futuro, gerado por um cliente durante o seu ciclo de vida junto à empresa.(Gold, 2020)

Na tentativa de minimizar o fenômeno "churn", as empresas focam na gestão de relacionamentos com consumidores com o objetivo de retê-los e prolongar ao máximo a relação dos clientes com a marca, principalmente no que diz respeito aos clientes mais rentáveis.

Com a disponibilização de tecnologia, modelos de aprendizado de máquina foram desenvolvidos permitindo a distinção de clientes, através do processo de segmentação, de forma que as empresas possam configurar inúmeras variações do mesmo produto/serviço com personalização de preços e recomendações de ofertas digitais. Além da customização em massa, isso possibilitou ações de retenção daqueles clientes mais interessantes para a organização, do ponto de vista de rentabilidade e lucratividade. (Gold, 2020)

Considerando que as perdas de lucro devido ao abandono de clientes podem ser significativas, a previsão de "churn" (abandono) representa informação relevante a gestores de negócios, pois clientes com altas probabilidades de evasão poderiam ser alvo de ações específicas de marketing, visando a sua retenção.

A inteligência artificial utiliza tecnologias de aprendizado de máquina como modelos de classificação, regressão logística, árvores de decisão e diversas técnicas estatísticas para investigar grandes volumes de dados e construir modelos que visam a predizer o comportamento dos clientes e da probabilidade de evasão deles, ou seja, quando ocorrerá o "churn" ou quando irão parar de consumir um produto/serviço.

O projeto pode auxiliar as empresas a compreenderem a taxa de abandono de seus clientes. A partir disso, os gestores do negócio podem personalizar ações de comunicação de marketing e tratamento diferenciado dos clientes segmentados conforme o risco de abandono, otimizando, assim, tempo, esforços e investimentos mercadológicos.

O objetivo desse projeto é aplicar modelos de "Machine Learning" para prever a evasão de clientes e apresentar um modelo de estimação do "churn" de clientes em relação a uma empresa, utilizando dados históricos de consumo e comportamento de clientes e descrever a validação do modelo e variáveis que influenciam no abandono ou na retenção.

## Material e Métodos

Para a elaboração deste estudo, utilizou-se a metodologia CRISP-DM ("Cross Industry Standard Process for Data Mining"), que consiste em um conjunto de boas práticas para se executar um projeto em Ciência de Dados.

A ideia dessa metodologia é mostrar que o processo de execução do projeto de Ciência de Dados não ocorre de maneira sequencial, mas sim de forma cíclica, onde o retorno a um estágio pode ocorrer, a fim de chegar em um resultado melhor e sempre focado no entendimento do negócio.

Esse processo, ilustrado na figura 1, é subdividido em 6 etapas, descritas a seguir.

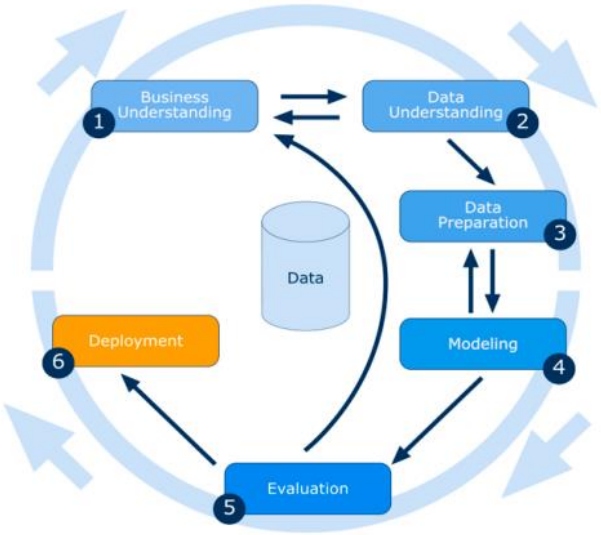


Figura1 - Fases do processo CRISP-DM para mineração de dados.
Fonte: Wirth e Hipp, 2000.

1. “Business Understanding” (Entendimento do negócio): Fase inicial focada no entendimento dos objetivos e requisitos do projeto, traçando um plano preliminar para que tais objetivos sejam alcançados;
2. “Data Understanding” (Entendimento dos dados): A coleta inicial de dados seguida de atividades de exploração e conhecimento deles, com a finalidade de identificar problemas na qualidade dos dados, descobrir insights ou reconhecer padrões refletidos no conjunto;
3. “Data Preparation” (Preparação dos dados): A fase de preparação de dados inclui seleção de variáveis, registro de atributos, limpeza de dados, construção de novos atributos e transformação de dados para ferramentas de modelagem;
4. “Modeling” (Modelagem): Nesta fase, técnicas de modelagem são selecionadas e aplicadas, e seus parâmetros são calibrados para valores ideais. Frequentemente, percebe-se problemas de dados durante a modelagem ou obtém-se ideias para a construção de novos dados, sendo necessário revisitar a fase de preparação;

5. “Evaluation” (Avaliação): Avaliar o modelo de forma detalhada e revisar as etapas executadas durante sua construção é de vital importância para garantir que seus resultados sejam adequados aos objetivos de negócios;
6. “Deployment” (Implementação): O novo conhecimento adquirido precisará ser organizado e apresentado de forma que seja possível aplicá-lo e usá-lo em problemas reais, a fim de fornecer uma resposta ao objetivo traçado durante o entendimento do negócio.

Os dados utilizados neste estudo procederam de um conjunto de dados públicos referente a pedidos enviados a uma loja brasileira de comércio eletrônico (“Olist Store”), e tem como metodologia o tipo qualitativo explicativo.

O foco da análise são as ordens de pedidos realizados na loja eletrônica, advindos da coleta dos dados, que se deu por meio da plataforma online de aprendizagem e competições para cientistas de dados conhecida como Kaggle.

O conjunto de dados possui informações de pedidos no período de 2016 a 2018 feitos em vários “marketplaces” no Brasil. Trata-se de dados reais, anonimizados e exportados em formato CSV.

A ferramenta utilizada para a importação, exploração, pré-processamento, modelagem e avaliação do conjunto de dados foi o Jupyter Notebook. O Jupyter Notebook é uma aplicação web que permite a criação e compartilhamento de código e textos.

A linguagem de programação utilizada no código foi a Python. As bibliotecas da linguagem de programação Python utilizadas para o projeto foram: Pandas, Numpy, Seaborn, Matplotlib, Datetime, Lifetimes, Yellowbrick e Sklearn.

**Descrição da Base de Dados**

O conjunto de dados públicos disponibilizados está organizado em 6 bases de dados distintas conforme a tabela 1.

Tabela 1. Bases de dados disponibilizadas

| Nome base de dados | Descrição |
|---|---|
| olist_order_payments_dataset | Dados referente aos pagamentos dos pedidos |
| olist_orders_dataset | Dados referentes aos pedidos |
| olist_customers_dataset | Dados referentes aos clientes |
| olist_order_reviews_dataset | Dados referente às avaliações feitas pelos clientes |
| olist_order_items_dataset | Dados referentes aos itens dos pedidos |
| olist_products_dataset | Dados referentes aos produtos pedidos |

Fonte: Dados originais da pesquisa

Tabela 2. Variáveis das bases de dados ‘olist_order_payments_dataset’

| Variáveis | Descrição | Tipo Variável |
|---|---|---|
| order_id | Código identificação do pedido | Nominal |
| payment_sequential | Sequência pagamentos | Discreta |
| payment_type | Tipo de pagamento | Nominal |
| payment_installments | Parcelas de pagamento | Discreta |
| payment_value | Valor pagamento | Contínua |

Fonte: Dados originais da pesquisa

A base de dados ‘olist_order_payments_dataset’ apresenta 5 variáveis (2 categóricas e 3 numéricas) e 103.886 observações.

Tabela 3. Variáveis das bases de dados ‘olist_orders_dataset’

| Variáveis | Descrição | Tipo Variável |
|---|---|---|
| order_id | Código identificação do pedido | Nominal |
| customer_id | Código identificação do cliente | Nominal |
| order_status | Status do pedido | Nominal |
| order_purchase_timestamp | Data do pedido de compra | Nominal |
| order_approved_at | Pedido aprovado em | Nominal |
| order_delivered_carrier_date | Data pedido entregue a transportadora | Nominal |
| order_delivered_customer_date | Data pedido entregue ao cliente | Nominal |
| order_estimated_delivery_date | Data estimada da entrega do pedido | Nominal |

Fonte: Dados originais da pesquisa

A base de dados ‘olist_orders_dataset’ apresenta 8 variáveis (todas categóricas) e 99.441 observações.

Tabela 4. Variáveis das bases de dados ‘olist_customers_dataset’

| Variáveis | Descrição | Tipo Variável |
|---|---|---|
| customer_id | Código identificação do cliente | Nominal |
| customer_unique_id | Código único do cliente | Nominal |
| customer_zip_code_prefix | Prefixo do CEP do cliente | Discreta |
| customer_city | Cidade do cliente | Nominal |
| customer_state | Estado do cliente | Nominal |

Fonte: Dados originais da pesquisa

A base de dados ‘olist_customers_dataset’ apresenta 5 variáveis (4 categóricas e 1 numérica) e 99.441 observações.

Tabela 5. Variáveis das bases de dados ‘olist_order_reviews_dataset’

| Variáveis | Descrição | Tipo Variável |
|---|---|---|
| review_id | Código identificação avaliação | Nominal |
| order_id | Código identificação do pedido | Nominal |
| review_score | Pontuação da avaliação | Discreta |
| review_comment_title | Título do comentário da avaliação | Nominal |
| review_comment_message | Mensagem do comentário da avaliação | Nominal |
| review_creation_date | Data criação da avaliação | Nominal |
| review_answer_timestamp | Data respostas da avaliação | Nominal |

Fonte: Dados originais da pesquisa

A base de dados ‘olist_order_reviews_dataset’ apresenta 7 variáveis (6 categóricas e 1 numérica) e 103.387 observações.

Tabela 6. Variáveis das bases de dados ‘olist_order_items_dataset’

| Variáveis | Descrição | Tipo Variável |
|---|---|---|
| order_id | Código identificação do pedido | Nominal |
| order_item_id | Código item do pedido | Discreta |
| product_id | Código identificação produto | Nominal |
| seller_id | Código identificação vendedor | Nominal |
| shipping_limit_date | Data limite do envio | Nominal |
| price | Preço | Contínua |
| freight_value | Valor do frete | Contínua |

Fonte: Dados originais da pesquisa

A base de dados ‘‘olist_order_items_dataset’ apresenta 7 variáveis (4 categóricas e 3 numéricas) e 112.650 observações.

Tabela 7. Variáveis das bases de dados ‘olist_products_dataset’

| Variáveis | Descrição | Tipo Variável |
|---|---|---|
| product_id | Código identificação do produto | Nominal |
| product_category_name | Nome categoria do produto | Nominal |
| product_name_lenght | Comprimento nome do produto | Contínua |
| product_description_lenght | Comprimento descrição do produto | Contínua |
| product_photos_qty | Quantidade fotos do produto | Contínua |
| product_weight_g | Peso em g do produto | Contínua |
| product_length_cm | Comprimento do produto em cm | Contínua |
| product_height_cm | Altura do produto em cm | Contínua |
| product_width_cm | Largura do produto em cm | Contínua |

Fonte: Dados originais da pesquisa

A base de dados ‘olist_products_dataset’ apresenta 9 variáveis (2 categóricas e 7 numéricas) e 32.951 observações.

## Pré-Processamento

Alguns pré-processamentos foram realizados para a análise exploratória dos dados:

- Ajuste da variável ‘order_purchase_timestamp’ para tipo data
- Adicionada à base ‘olist_orders_dataset’ variáveis do dia, mês e ano da ordem de pedido
- Identificada a data da primeira observação e última observação do pedido de compra da base ‘olist_orders_dataset’
- Adicionada variável ‘total_sales’ na base ‘olist_order_items_dataset’ referente à soma do valor do pedido e valor do frete

## Descrição das Variáveis

### Pedidos de Compra

Na análise exploratória dos dados, 0,3% dos pedidos foram realizados no ano de 2016, enquanto 45,4% aconteceram em 2017 e 54,3% ocorreram em 2018. Observa-se também, que há uma evolução positiva nas vendas entre o período de setembro de 2016 a agosto de 2018.

### Análise Clientes

Na análise das vendas por estado brasileiro, o maior volume de vendas está concentrado no estado de São Paulo, com 37% do total, seguido por Rio de Janeiro com 13% do total.

### Análise Tipos de Pagamento

Os tipos de pagamentos utilizados nos pedidos foram: cartão de crédito, cartão de débito, boleto e voucher. A maior representatividade de pagamento com 73,9% foi o cartão de crédito, seguido pelo pagamento por boleto com 19,0%, voucher com 5,6% e cartão de débito com 1,5%.

### Análise Avaliação de Clientes

Na avaliação dos clientes com pontuação de 1 a 5, onde 1 representa os clientes insatisfeitos e 5 os satisfeitos, 56,5% dos clientes estão satisfeitos e avaliaram o negócio com a pontuação 5, seguidos por 19% (pontuação 4), 8,4% (pontuação 3), 3,4% (pontuação 2) e 12,7% (pontuação 1).

### Definição Variável Alvo (target)

Como explicado acima, o abandono ou “churn” é um termo de Marketing utilizado para identificar um cliente que vai abandonar a empresa. E, para se antecipar a essa ocorrência e traçar estratégias efetivas, o primeiro passo é identificar os clientes que irão abandonar a empresa.

O “churn” está diretamente associado ao tempo em que o indivíduo permanece como cliente de uma organização, portanto está diretamente relacionado ao conceito de “Consumer Life Time Value” (CLV ou valor do ciclo de vida do cliente).

Neste estudo a maneira de classificar um cliente como “churn” ou “non-churn” foi feita através do “Customer Lifetime Value“ ou da análise RFM (Recência, Frequência, Monetária).

A análise RFM foi usada para a criação da variável alvo (target) que é a variável que irá caracterizar se o cliente fará a evasão ou não (“churn” ou “não-churn”).

Adicionada a análise RFM, aplicou-se o modelo PCA – Análise dos Componentes Principais para a redução de dimensionalidade da base de dados e então aplicou-se o modelo de clusterização ou agrupamento K-means com o objetivo de reconhecer os padrões dos clientes quanto ao abandono do negócio e segmentá-los, de formar a propor estratégias de comunicação e/ou novos produtos/serviços dirigida a eles.

### Modelo Customer Lifetime Value

O principal objetivo deste estudo é construir um modelo preditivo que permita estimar o CLV futuro de um cliente e assim em última análise aumentar o lucro da organização, dado que o CLV permite um melhor conhecimento do cliente e que se desenhem campanhas personalizadas para cada cliente. Com esse propósito, define-se como objetivo específico a implementação de modelos “Machine Learning”. Define-se também como objetivo deste trabalho a realização de uma análise de clusters, com o objetivo de inferir quais são as variáveis que contribuem de forma mais significativa para o aumento ou diminuição do CLV de determinado grupo de clientes.

O valor vitalício do cliente (CLV) é o valor total de um cliente para uma empresa ao longo de seu relacionamento. Na prática, esse “valor” pode ser definido como receita, lucro ou outras métricas de escolha do analista.

O CLV é uma métrica importante a ser rastreada por dois motivos. Primeiro, a totalidade do CLV de uma empresa sobre toda a sua base de clientes dá uma ideia aproximada de seu valor de mercado. Em segundo lugar, uma análise de CLV pode orientar a formulação de estratégias de aquisição e retenção de clientes.

O Customer Lifetime Value (CLV) é um conceito de gestão de clientes definido há mais de 30 anos por Kother como o valor presente do fluxo de lucro futuro esperado num determinado horizonte temporal de transações com o cliente.

A segmentação por Recência, Frequência e Valor Monetário (RFM) utiliza um método para fazer a distinção por meio da identificação dos clientes levando em conta a última vez que o consumidor fez uma compra (recência), qual é a frequência das aquisições feitas (frequência) e a quantia gasta (valor monetário). Nos últimos anos tem-se optado por usar as variáveis R, F e M como input para modelos de “Machine Learning”.

A previsão do valor da vida útil do cliente pode ser calculada utilizando os modelos Beta Geometric Negative Binomial Distribution (BG/NBD) e o submodelo Gama-Gama.

O modelo de Distribuição Beta Geométrica / Binomial Negativa (BG/NBD), modela a distribuição dos comportamentos de compra de cada cliente e prevê o número esperado de transações para cada cliente, e seu submodelo Gamma-Gamma modela a distribuição de lucro médio esperado e prevê o lucro médio esperado para cada cliente.

O modelo BG-NBD aborda situações de compras realizadas sem contrato/obrigação e com ocorrência a qualquer momento, ou seja, sem regularidade/periodicidade

Sob tal configuração, a evasão do cliente não é explicitamente observável e pode acontecer a qualquer momento. Isso torna mais difícil diferenciar clientes que se desligaram indefinidamente daqueles que retornarão no futuro. O modelo BG/NBD é capaz de atribuir probabilidades a cada uma dessas opções.

O modelo BG/NBD prevê o número de transações de um determinado cliente em um período através do cálculo de número de compras multiplicado pelos valores das compras.

O submodelo Gamma-Gamma prevê o valor esperado de compras de um determinado cliente em um período através do cálculo da média de todas as compras anteriores.

O modelo BG/NBD é um modelo probabilístico e assume que as observações (ou seja, as transações) são geradas por um processo físico que pode ser modelado usando distribuições de probabilidade.

A seguir, algumas as propriedades do modelo BG/NBD:

- Quando um usuário está ativo, um número de suas transações em um período de tempo t é descrito pela distribuição de Poisson com taxa de transação λ.
- A heterogeneidade na taxa de transação entre os usuários (como os clientes diferem no comportamento de compra) tem distribuição Gama com parâmetros r (tamanho) e α (escala).
- Os usuários podem ficar inativos (“churn”) após qualquer transação com probabilidade p e seu ponto de desistência (quando ficam inativos) é distribuído entre compras com distribuição Geométrica.
- A heterogeneidade (variação entre usuários) na probabilidade de abandono tem distribuição Beta com os dois parâmetros de forma α e β.
- A taxa de transação e a probabilidade de abandono variam independentemente entre os usuários.

Notação matemática para representar as características de um usuário X:

X = x, t_x, T, onde x é o número de transações em algum período de tempo (0, T], e t_x (0 < tx ≤ T) é o período da última compra.

Com base nesses parâmetros, o modelo prevê futuros padrões de compra dos clientes:

P(X(t) = x) — probabilidade de observar x transações no período t no futuro

E(Y(t) | X = x, t_x, T) — número esperado de transações no período para um cliente com comportamento observado.

A partir dessas características é possível obter a probabilidade do cliente estar ativo (no-”churn”),

$$P(X(t)=x|\lambda,p)=(1-p)x\frac{(\lambda t)^x e^{-\lambda t}}{x!}+\ \delta x>0p(1-p)^{x-1}\left[1-e^{-\lambda t}\sum_{j=0}^{x-1}\frac{(\lambda t)^j}{j!}\right], \quad (1)$$

$$Onde, \delta x>0=1\ se\ x>0\ , caso\ contrário\ 0$$

E prever o número de transações

$$\mathrm{E}(Y(y)|X=x,t_x,T,r,\alpha,a,b)=\frac{\frac{a+b+x-1}{a-1}\left[1-\left(\frac{\alpha+T}{\alpha+T=t}\right)^{r+x}{}_2F1(r+x,b+x;a+b+x-1;\frac{t}{\alpha+T+t}\right]}{1+\delta x>0\ \frac{a}{b+x-1}(\frac{\alpha+T}{\alpha+t_x})^{r+x}} \quad (2)$$

$$Onde\ 2F1\ \text{é}\ a\ função\ hipergeométrica\ gaussiana$$

$$2F1(a,b;c;z)=\frac{\Gamma(c)}{\Gamma(a)\Gamma(b)}\sum_{j=0}^{\infty}\frac{\Gamma(a+j)\Gamma(b+j)}{\Gamma(c+j)}\frac{z^j}{j!} \quad (3)$$

Em suma, obtendo as estimativas dos parâmetros do modelo r, α, a, b (por exemplo, usando o estimador de máxima verossimilhança), é possível prever o número esperado de transações para os usuários.

Após a aplicação do modelo BG/NBD para obter a recência, frequência e valor monetário e prever o número de transações dos clientes, é possível prever os valores futuros das transações com o submodelo Gamma-Gamma e, com todos os elementos, determinar o valor de vida útil de cada cliente.

CLV = número esperado de transações * receita por transação * margem

Onde o primeiro elemento refere-se ao modelo BG/NBD, o segundo elemento ao modelo Gamma-Gamma e a margem é definida pelo negócio.

Notação matemática para o modelo Gamma-Gamma:

O cliente tem x transações com valores z1, z2, …, m_x = Zi/x é observado o valor médio da transação

E(M) é o valor médio não observado da transação e E(M | mx, x) é o valor monetário esperado de um cliente de acordo com seu comportamento de compra.

As propriedades do modelo Gamma-Gamma são:

- O valor monetário das transações dos usuários é aleatório em torno do valor médio da transação.
- O valor médio da transação varia entre os usuários, mas não varia para um usuário individual ao longo do tempo.
- Os valores médios das transações têm distribuição Gamma entre os clientes.

Assim o modelo Gamma se apresenta

$$\mathrm{E}(\mathrm{M}|p,q,\gamma,m_x,x) = \frac{(\gamma+m_x x)p}{px+q-1} = \left(\frac{q-1}{px+q-1}\right)\frac{\gamma p}{q-1} + \left(\frac{px}{px+q-1}\right) m_x \ , \qquad (4)$$

onde p é o tamanho e v são os parâmetros de escala da distribuição gama para transações Zi, q é o tamanho e γ são os parâmetros de escala para a distribuição gama de v (p é constante por suposição - o coeficiente de variação de nível individual é o mesmo para todos os clientes). Utiliza-se o método de máxima verossimilhança para estimar os parâmetros do modelo.

**Modelo de PCA – Principal Components Analysis**

Essa técnica é muito utilizada para a redução de dimensionalidade por ser uma técnica não paramétrica (não possui parâmetros) e ser linear, simples e rápida.

O PCA é baseado na variância dos dados, ou seja, cria uma nova representação dos dados, com uma dimensão menor, mantendo a variância entre eles. Seus novos eixos são

descorrelacionados. Cada eixo possui uma variância, normalmente dada em % em relação ao todo.

Neste estudo utilizou-se o modelo PCA para a redução de dimensionalidade da base de dados como entrada para o modelo de clusterização K-means.

Ao manter um número grande de variáveis em um conjunto de dados, a flexibilidade no processo de modelagem é prejudicada e as análises se tornam mais complexas. Isso ocorre porque o número de modelos aplicáveis à base de dados com essas características é mais restrito do que aquele aplicável à bases de dados mais enxutas. Outro ponto a ser levantado nesse aspecto é o espalhamento dos dados. Quando uma base de dados possui muitas variáveis, torna-se mais difícil clusterizar os dados e isso ocorre porque, aparentemente, os pontos ficam aproximadamente equidistantes entre si.

Como na clusterização alguma medida de distância é utilizada para quantificar a similaridade entre as observações, como, por exemplo, a distância euclidiana, isso configura-se como um grande problema. Se as distâncias entre os pontos são aproximadamente iguais, nenhum grupo significativo pode ser formado. Além disso, quando há mais de três dimensões, a visualização dos dados se torna muito mais difícil para observar o comportamento destes quando analisados conjuntamente.

Além dos pontos colocados acima, a redução de dimensionalidade é relevante para o overfitting. Esse fenômeno ocorre quando um modelo se ajusta demasiadamente aos dados de um conjunto de treinamento, capturando, até mesmo, o ruído desse conjunto e apresentando um desempenho inferior no conjunto de testes. À medida que o número de variáveis aumenta, o modelo se torna cada vez mais dependente dessas variáveis, se “perdendo” nas entrelinhas de dados não tão relevantes para o resultado final que se almeja alcançar, de modo que corre o risco de ele deixar de capturar apenas a tendência dos dados para capturar também o ruído destes. A performance do modelo no conjunto de testes, nessa condição, diminui em vez de aumentar, indo na contramão daquilo que se objetiva ao preservar um número maior de dimensões no conjunto de dados.

De modo geral, empregar algoritmos de redução de dimensionalidade em base de dados elegíveis para tal auxilia o processo de modelagem desses dados, bem como pode trazer aumento de performance aos modelos obtidos. Quando reduzido o número de variáveis, eliminando a redundância nos dados — isto é, variáveis distintas que significam a mesma coisa — e mantendo apenas a tendência do conjunto, a qual pode ser visualizada facilmente em duas ou três dimensões, é possível aplicar uma quantidade superior de modelos a esses dados, processá-los rapidamente e, assim, realizar testes e ajustes de hiperparâmetros iterativamente para aumentar a performance e a qualidade dos resultados. Além disso, os modelos ficam menos propensos a overfitting e os resultados são mais fáceis

de visualizar, o que torna todo o processo de modelagem mais ágil. Por fim, o custo computacional de armazenamento também cai bastante, haja vista que o conjunto se torna mais compacto.

Quando aplicado o método PCA, basicamente, a matriz de variáveis correlacionadas é substituída por um conjunto novo de variáveis não-correlacionadas, dotadas de alta variância. Se eliminadas as variáveis correlacionadas, elimina-se a redundância da base de dados, retendo apenas o que é relevante, e se priorizada a variância elevada, prioriza-se também o alto número de informações.

Sendo assim, quando reduzida a dimensionalidade de um conjunto de dados e projetada um elevado número de variáveis em um número reduzido de dimensões, deve-se procurar manter a variância dos dados, porque assim perde-se o menor número de informações possível.

**Modelo Clusterização K-means**

Clustering ou análise de cluster é uma técnica que permite encontrar grupos de objetos semelhantes, objetos que estão mais relacionados entre si do que com objetos em outros grupos. O conjunto de dados semelhantes são chamados grupos ou clusters. A palavra semelhante entende-se como que possuem características próximas.

O K-means clustering faz parte dos algoritmos pertencentes ao aprendizado não supervisionado que é o ramo do aprendizado de máquina (“Machine Learning”) que aprende a partir de dados que não foram rotulados, classificados ou categorizados. Em vez de responder ao feedback, a aprendizagem não supervisionada identifica as semelhanças nos dados e reage com base na presença ou ausência dessas semelhanças em cada novo dado.

No agrupamento de K-means o algoritmo separa os dados em K clusters a partir de dados que devem estar em forma de vetores numéricos. O método calcula a média de um conjunto de pontos de dados e a distância euclidiana entre eles.

O algoritmo K-means pertence à família de algoritmos chamados de algoritmos de otimização de agrupamento. Ou seja, os exemplos são divididos em grupos de clusters, de forma que o cluster dê bons resultados de acordo com os critérios definidos. O nome do algoritmo foi derivado de forma que os K clusters são formados a partir dos conjuntos de dados em que o centro do cluster é a média aritmética de todos os objetos dentro desse tipo de cluster. O número de K clusters é conhecido de antemão. A primeira etapa é encontrar os centróides iniciais para cada cluster. A próxima etapa é associar cada objeto de dados ao seu centróidde mais próximo. O agrupamento inicial é feito atribuindo a cada dado, objeto ao centróide que está tão próximo a ele e a primeira iteração é concluída. O algoritmo funciona

em iterações até que os objetos não mudem seus centros de cluster. Os centróides movem suas posições até que os critérios de convergência sejam alcançados.

Ressalta-se que a base do modelo K-means é que o centro do cluster (centróide) é a média aritmética de todos os pontos pertencentes ao cluster e cada ponto está mais próximo de seu próprio centróide do que outros centróides.

O algoritmo k-means requer a especificação do número de clusters, que pode ser orientado pela necessidade da entidade que a aplica. Na ausência de um número de cluster ditado, uma abordagem estatística pode ser usada. Uma abordagem comum, chamada de método do cotovelo (Elbow method), identifica quando o conjunto de cluster explica a maior parte da variação nos dados O cotovelo é o ponto onde a variância cumulativa explicada se aplaina depois de subir abruptamente, daí o nome do método.

**Resultados Preliminares**

Após a fusão das bases de dados 'olist_orders_dataset', 'olist_customers_dataset' e ‘olist_order_items_dataset’ aplicou-se a função summary_data_from_transaction_data do modelo BG/NBD para obter a frequencia, recencia e tempo e valor monetário de cada cliente, conforme tabela a seguir:

Tabela 8. Resumo variáveis frequência, recência, tempo e valor monetário

| | Frequency | Recency | T | Monetary_value |
|---|---|---|---|---|
| count | 95420.000000 | 95420.000000 | 95420.000000 | 95420.000000 |
| mean | 0.024198 | 2.683389 | 246.184846 | 2.474897 |
| std | 0.178936 | 25.263074 | 153.656402 | 26.579606 |
| min | 0.000000 | 0.000000 | 5.000000 | 0.000000 |
| 25% | 0.000000 | 0.000000 | 122.000000 | 0.000000 |
| 50% | 0.000000 | 0.000000 | 227.000000 | 0.000000 |
| 75% | 0.000000 | 0.000000 | 356.000000 | 0.000000 |
| max | 15.000000 | 633.000000 | 729.000000 | 1999.990000 |

Fonte: Resultados originais da pesquisa

Após aplicar a condição de frequência maior que 0 na base, foi aplicada a função BetaGeoFitter do modelo BG/NBD com coeficiente de penalização 0.00 para as variáveis

frequência, recência e tempo gerando “fit” do modelo com as propriedades conforme imagem abaixo

Tabela 9. Propriedades do modelo BG/NBD BetaGeoFitter

| | Coef | se(coef) | lower 95% bound | upper 95% bound |
|---|---|---|---|---|
| r | 0.015627 | 0.000819 | 0.014021 | 0.017233 |
| alpha | 65.858911 | 6.194610 | 53.717476 | 78.000347 |
| a | 2.068727 | 0.494811 | 1.098898 | 3.038556 |
| b | 0.353572 | 0.092534 | 0.172206 | 0.534938 |

Fonte: Resultados originais da pesquisa

Através da função de visualização da matriz Frequência/Recência a partir do BetaGeoFitter, que calcula o número esperado de transações que um cliente fará no próximo período de tempo, considerando sua recência (idade na última compra) e frequência (o número de transações repetidas ele ou ela fez), é possível observar que o melhor cliente comprou 15 vezes no período mais recente (700 dias).

Para avaliar se o modelo está performando de forma correta aplicou-se a função plot_period_transactions no modelo BetaGeoFitter que compara os dados inseridos no modelo com os dados previstos do modelo ajustado.

Os dados imputados no modelo e os dados previstos estão alinhados em até 3 repetições e que a partir de 4, há poucos clientes que realizaram compra. Pelo alinhamento dos dados identifica-se uma boa performance no modelo.

Utilizando a função conditional_probability_alive do modelo BG/NBD define-se um limite para clientes que já desistiram (“churn”) e os que correm o risco de desistir, mas ainda não desapareceram. A partir dessa função identifica-se que clientes abaixo de 0.1 (10%) são “churn” (desistentes) enquanto os clientes acima de 0.1 e abaixo de 0.2 considera-se alto risco de desistência e os acima de 0.2 considera-se não desistentes (non-”churn”).

Para as previsões de compras para o período de 6 meses futuros, aplicou-se a função conditional_expected_number_of_purchases_up_to_time do modelo BG/NBD Beta Geo Fitter

Para o cálculo do Life Time Value (valor do ciclo de vida do cliente) aplicou-se o submodelo GammaGammaFitter do BG/NBD com coeficiente de penalização 0.00 para as variáveis frequencia e valor monetário gerando “fit” do modelo com as propriedades conforme imagem abaixo

Tabela 10. Propriedades do modelo BG/NBD GammaGammaFitter

| | coef | se(coef) | lower 95% bound | upper 95% bound |
|---|---|---|---|---|
| p | 2.977437 | 0.226051 | 2.534378 | 3.420496 |
| q | 2.900954 | 0.204591 | 2.499955 | 3.301952 |
| v | 72.164356 | 11.407650 | 49.805361 | 94.523350 |

Fonte: Resultados originais da pesquisa

Para as previsões do ciclo de vida dos clientes para os próximos 6 meses, aplicou-se a função customer_lifetime_value do submodelo GammaGamma do BG/NBD.

No quadro a seguir é demonstrado o sumário do estudo contendo as variáveis de frequencia, recencia, tempo, valor monetário, probabilidade de “churn”, compras futuras e ciclo de vida futuro.

Tabela 11. Resumo dos valores de frequência, recência, tempo, valor monetário, probabilidade de “churn”, compras futuras e LTV resultante do modelo BG/NBD

| | **frequency** | **recency** | **T** | **monetary_value** | **prob_alive** | **purchase_next_6_month** | **LTV_next_6_month** |
|---|---|---|---|---|---|---|---|
| count | 2085 | 2085 | 2085 | 2085 | 2085 | 2085 | 2085 |
| mean | 1.107 | 122.805 | 319.305 | 113.264 | 0.061 | 0.035 | 13.294 |
| std | 0.516 | 120.263 | 146.804 | 140.686 | 0.058 | 0.087 | 27.771 |
| min | 1.000 | 1.000 | 10.000 | 3.850 | 0.005 | 0.001 | 0.555 |
| 25% | 1.000 | 25.000 | 202.000 | 40.900 | 0.028 | 0.011 | 4.138 |
| 50% | 1.000 | 80.000 | 321.000 | 78.000 | 0.049 | 0.020 | 7.718 |
| 75% | 1.000 | 188.000 | 438.000 | 135.910 | 0.077 | 0.034 | 13.116 |
| max | 15.000 | 633.000 | 700.000 | 1999.990 | 0.800 | 3.045 | 646.779 |

Fonte: Resultados originais da pesquisa

A partir da base de dados elaborada com as variáveis citadas acima, aplicou-se o modelo PCA – Análise dos Componentes Principais para a redução de dimensionalidade da base e obtenção de 2 componentes baseados na variância dos dados para servirem de entrada no modelo de clusterização aplicado a seguir:

No modelo de clusterização utilizou-se o método Elbow para definição da quantidade de k clusters. Com o número de 3 grupos obtidos pelo método do cotovelo foi possível inserir o parâmetro no modelo K-means para treinar o modelo e identificar os centróides de cada um dos grupos

Com os centróides e os rótulos (‘*labels*’) de cada grupo (cluster) definidos, foi adicionada à base de dados uma nova variável com os clusters do modelo K-means para identificação de perfil de cada cliente.

A seguir as características dos clientes identificados em cada um dos 3 grupos:

Grupo 0 – representam 11% do volume dos clientes. Apresentam alta frequência e recência de compras, apresentam um valor monetário médio, são clientes de non-”churn” (não desistentes) e alto risco de “churn”, possuem um volume de compras médio para os próximos seis meses e um altíssimo valor de LTV.

Grupo 1 – representam 41% do volume dos clientes. Apresentam média frequência e recência de compras, apresentam um valor monetário baixo, são clientes desistentes (“churn”), possuem um valor de compras baixo para os próximos seis meses e um baixo valor de LTV

Grupo 2 – representam 48% do volume dos clientes. Apresentam baixa frequência e recência de compras, apresentam um valor monetário alto, são clientes de “churn” e (desistentes) e alto risco de “churn”, possuem um valor de compras alto para os próximos seis meses e um médio valor de LTV.

**Conclusão**

Ao longo do estudo foi apresentado o comportamento de consumo e propensão de evasão / “churn” dos clientes em relação à empresa.

Aproximadamente 89% dos clientes desistiram (“churn”), o que significa que há muitas oportunidades de melhoria em relação à retenção. Pode-se supor que os clientes que desistiram já estão perdidos, porém ainda existem os clientes que correm alto risco de “churn”, mas que ainda não desistiram. Com foco nesses clientes (presentes nos clusters 0 e 2), é possível realizar ações de Marketing como: comunicações personalizadas, promoções, programas de fidelização para tratamentos direcionados de modo a conquistá-los e evitar o “churn”.

A utilização dos modelos de aprendizado de máquina e as análises estatísticas que descreveram o comportamento e padrões dos clientes, viabilizou a segmentação e

identificação de perfil dos clientes, de forma a contribuir com o suporte para estratégias e ações para aumento de retenção, minimização de evasão, aumento de receitas, melhorias de produtos/serviços, marketing/comunicação dirigida consequentemente aumentando a competitividade da empresa no mercado.

## Agradecimentos

À professora Daniele Aparecida Cicillini Pimenta, pela orientação e oportunidade de crescimento acadêmico.

Ao Paulo Cotta, colega de trabalho que acreditou na minha transformação e me forneceu o suporte para os novos desafios no desenvolvimento da carreira na área de Ciência de Dados.

À minha companheira Lis de Oliveira pelo incentivo, suporte, lições de vida, paciência e carinho.

## Referências

Bussab, Wilton de Oliveira , Morettin, Pedro Alberto. Estatística básica. Editora Saraiva, São Paulo, SP, Brasil.

Fávero, Luís Paulo, Belfiore, Patrícia. 2019. Data Science for Business and Decision Making, Editora Academic Press, London, United Kingdom

Fávero, Luís Paulo, Belfiore, Patrícia.2017. Manual de Análise de Dados - Estatística e Modelagem Multivariada com Excel®, SPSS® e Stata, Editora GEN LTC, Rio de Janeiro, RJ, Brasil.

Fawcett, Tom, 2013.Data Science para Negócios: o que você precisa saber sobre mineração de dados e pensamento analítico de dados. 1ed. Editora Alta Books, Rio de Janeiro, RJ, Brasil.

Gold, Carl.2020. Fighting “churn” with Data: The Science and Strategy of Customer Retention. Editora Manning Publications, Shelter Island, New York, United States.

Grus, Joel. 2016. Data science do zero: primeiras regras com o Python, Editora Alta Books, Rio de Janeiro, RJ, Brasil.

Kotler, Philip, Kartajaya, Hermawan, Setiawan Iwan, 2017. Marketing 4.0: do tradicional ao digital, Editora Sextante, Rio de Janeiro, RJ, Brasil.

Kotler, Philip, Kartajaya, Hermawan, Setiawan Iwan, 2021. Marketing 5.0: tecnologia para a humanidade, Editora Sextante, Rio de Janeiro, RJ, Brasil.

Magaldi, Sandro , Salibi Neto, José. 2018. Gestão do amanhã: Tudo o que você precisa saber sobre gestão, inovação e liderança para vencer na 4a Revolução Industrial, Editora Gente, São Paulo, SP, Brasil.

Mckinney, Wes 2018. Python Para Análise de Dados: Tratamento de Dados com Pandas, NumPy e IPython, Editora Novatec, São Paulo, Sp, Brasil.

Noah Harari, Yuval. 2016. Homo Deus: uma breve história do amanhã. 1ed. Editora Companhia das Letras, São Paulo, SP, Brasil.

Schwab, Klaus. 2018. A Quarta Revolução Industrial. Tradução: Daniel Moreira Miranda. 1a ed. Editora Edipro, São Paulo, SP, Brasil

Wheelan,Charles. 2016.Estatística: O que é, para que serve, como funciona, Editora Zahar, Rio de Janeiro, RJ, Brasil.

Wirth, Ruediger, Hipp, Jochen. 2000. Crisp-dm: Towards a standard process model for data mining. Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining. London, United Kingdom.

Kaggle. Brazilian E-Commerce Public Dataset by Olist. Disponível em: <https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce?select=olist_sellers_dataset.csv > Acesso em: 10 abril de 2022.