Abordagem analítica para predição e prevenção do Churn

Analytical approach to churn identification and prevention

George Alves[1]
orcid.org/0000-0002-0170-8057

Lucas Lima[1]
orcid.org/0000-0002-7202-

Lucas Oliveira[1]
orcid.org/0000-0002-6320-

[1]Escola Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil. E-mail: gvsa@ecomp.poli.br

DOI: 10.25286/repa.v7i3.2461

Como citar este artigo pela NBR 6023/2018: George Alves; Lucas Lima; Lucas Oliveira. Abordagem analítica para predição e prevenção do Churn. Revista de Engenharia e Pesquisa Aplicada, Recife, v. 7, n. 3, p. 64-72.

RESUMO

O Churn, é um termo que se refere a clientes que abandonam uma empresa, este problema é constante no mundo empresarial. Dessa forma se torna necessário o uso de técnicas de análise e tratamento dos dados, para entender e solucionar o processo de Churn numa empresa. A empresa analisada nesta pesquisa foi a Justa, que é uma Fintech brasileira, que proporcionou a base de dados para avaliação e implementação deste estudo. A base disponibilizada contém duas partes: As informações dos clientes em si e as transações deles, nestas foram realizadas etapas de pré-processamento para melhor análise dos dados. Após as etapas de pré-processamento são aplicados técnicas e algoritmos de Machine Learning como: K-means, KNN e Logistic Regression a fim de buscar solucionar o problema de Churn na empresa. Os resultados aqui obtidos mostram que, para o escopo estimado, o projeto consegue dizer se um cliente é churn, com base nas suas transações, mas devido a grande rotatividade de clientes os grupos de clientes analisados não são acentuados e possuem poucos padrões comportamentais. Para uma análise mais elaborada dos perfis de cliente, é necessário obter informações mais detalhadas

PALAVRAS-CHAVE: Churn; Regressão; Clustering; Classificação;

ABSTACT

Churn is a term that refers to customers who leave a company, this problem is constant in the business world. Thus, it becomes necessary to use data analysis and processing techniques to understand and solve the Churn process in a company. The company analyzed in this research was Justa, which is a Brazilian Fintech company, which provided the database for the evaluation and implementation of this study. The available base contains two parts: The customer information itself and their transactions, in which pre-processing steps were carried out for better data analysis. After the pre-processing steps, Machine Learning techniques and algorithms such as: K-means, KNN and Logistic Regression are applied in order to seek to solve the Churn problem in the company. The results obtained here show that, for the estimated scope, the project is able to say whether a customer is churn, based on their transactions, but due to the high turnover of customers, the groups of customers analyzed are not accentuated and have few behavioral patterns. For a more elaborate analysis of customer profiles, it is necessary to obtain more detailed customer information, such as monthly income, occupancy, among others.

KEY-WORDS: Churn; Regression; Clustering; Classification;;

REVISTA DE Engenharia e Pesquisa Aplicada

1 INTRODUÇÃO

Com o progresso da tecnologia o mercado vem tornando-se cada vez mais competitivo e a manutenção da base de clientes se mostra de suma importância para qualquer negócio. Não basta ter muitos clientes e uma boa experiência de usuário, pois, com tantos competidores, as empresas precisam estar sempre inovando para aumentar e manter a base de clientes ativos.

A base de clientes de um negócio é composta pelos usuários ativos e inativos. Usuários ativos são aqueles que, dada uma janela de tempo, estão consumindo recursos ou serviços da empresa. Saber por que um cliente abandona ou deixa de comprar ou utilizar o seu produto ou serviço é fundamental para garantir um crescimento sustentável para a sua empresa [1].

De acordo com um estudo da Ipsos Loyalty, cerca de 20% dos clientes abandonam uma empresa durante o ano, esse evento é chamado de churn, ou seja, o processo de desistência do consumidor com relação ao produto ou serviço. Em linhas gerais, o churn é o caso em que os clientes param de realizar as atividades na empresa.

Existem diversas formas de analisar a retenção destes clientes, uma delas é através do churn rate (taxa de rotatividade) [2]. De acordo com o dicionário de Oxford, o churn rate significa: "the annual percentage rate at which customers stop subscribing to a service or employees leave a job" (Oxford, 2020). Em tradução livre: é a taxa percentual de consumidores que deixam de consumir ou empregar um serviço. Essa é uma medida que está intrinsecamente relacionada aos resultados de uma empresa, trata-se do dimensionamento da quantidade de usuários ou clientes que abandonaram os produtos ou serviços de uma determinada empresa [3].

O churn rate, normalmente, é calculado considerando uma janela de tempo. O percentual do churn é obtido através da divisão do número de clientes que abandonaram o serviço ou produto, pelo número total de clientes. Este valor também é importante para evidenciar o impacto financeiro desses cancelamentos, no caixa da empresa [2]. Compreender o churn rate é de grande importância para uma boa gestão financeira e dos clientes de um empreendimento.

Neste artigo, nós apresentamos uma análise que é responsável por encontrar variáveis correlacionadas que levaram os clientes a parar de realizar transações, desta forma, auxiliando a empresa na redução dos impactos causados pelo churn. Como base para a análise, serão utilizados os dados disponibilizados pela empresa Justa.

Este trabalho tem por objetivo realizar uma análise sobre a massa de dados e identificar as, eventuais, causas da saída dos usuários, bem como os parâmetros correlacionados. Por outro lado, não serão apresentadas soluções práticas para o problema, o objetivo é fornecer os insumos necessários para o debate de uma solução posterior.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 CHURN E CHURN RATE

O *churn* de clientes, ou rotatividade de clientes, refere-se ao número de clientes que uma empresa perde em um determinado período de tempo. De acordo com o livro *Fighting Churn With Data*, isto ocorre quando os consumidores deixam de utilizar ou passam a cancelar a subscrição de um produto **[2]**.

O primeiro objetivo de qualquer empresa que oferece produtos ou serviços é crescer, ou seja, a relação entre clientes entrantes e saintes deve ser positiva, de modo que o número de clientes esteja sempre crescente **[1]**. A taxa de rotatividade dos clientes (*churn rate*) é um valor percentual que representa a relação entre clientes entrantes e saintes, em um dado período de tempo **[4]**.

O *churn rate* é uma métrica muito importante, uma vez que uma alta taxa de desistência implica em um aumento no custo do cliente, pois, por muitas vezes o cliente não ficou tempo suficiente para equiparar o CAC (*customer acquisition cost,* do inglês, custo de aquisição de cliente).

O cálculo do *churn rate* é possível através da razão entre o número de clientes perdidos pelo total de clientes do período, multiplicado por 100 (para obter valores percentuais). Prioritariamente, deve-se determinar um período de tempo, o número de clientes no começo deste período e o número de clientes que saíram da base no final do período, em seguida dividir o número de clientes perdidos pelo número total de clientes, por fim, multiplicar a razão por 100 (cem) **[2]**. Abaixo a representação gráfica:

$$\frac{Clientes\ Perdidos}{Total\ de\ Clientes} * 100$$

DOI: 10.25286/repa.v7i3.2461

Em empresas de alto crescimento, o *churn* deve ser calculado de uma outra forma, uma vez que o número de clientes pode variar muito rápido. Para o cálculo desse número, é necessário incluir o valor do número de clientes no começo e no final do período **[2].**

$$\left(\frac{2 * Clientes\ Perdidos}{Clientes\ no\ inicio + Clientes\ no\ final}\right) * 100$$

Para negócios sazonais a fórmula também precisa ser adaptada para considerar a variação do número de clientes nos diferentes períodos de operação da empresa. O valor **T** representa o número total de clientes e o **P** representa o número de clientes perdidos **[2]**.

$$\left(\frac{T_{movimentado} * P_{movimentado} + T_{fraco} * P_{fraco}}{T_{movimentado} + T_{fraco}}\right) * 100$$

Reduzir o *churn rate* é equivalente a aumentar a retenção de clientes. Desta forma, faz-se necessário uma abordagem que consiga diminuir essa taxa, para evitar que os clientes deixem de utilizar os produtos da empresa **[1]**.

2.2 K-MEANS

Em 1967, o termo foi utilizado por James McQueen, porém, a ideia surgiu em 1957 através de Hugo Steinhaus. O agrupamento k-means é um método, sem supervisão, de separar, em torno de centros (centróides), diversos dados, criando os chamados clusters **[5]**.

O clustering é o conjunto de técnicas de análise de dados que consiste em fazer agrupamentos automáticos de dados segundo o seu grau de similaridade, no caso do k-means, este pretende particionar *n* observações dentre *k* grupos onde cada observação pertence ao grupo mais próximo da média **[6]**. Isso resulta em uma

divisão do espaço de dados em um Data Space constituído por células de Voronoi, que é um tipo especial de decomposição de Data Space.

O algoritmo consiste em definir um k, ou seja, um número de clusters (ou agrupamentos) e então aleatoriamente decidir um centróide para cada cluster definido. Com isto, pode-se calcular o centróide de menor distância para cada ponto, assim cada ponto pertencerá ao centróide mais próximo. Assim como o k pode ser definido, o distanciamento para o centróide também pode, assim as distâncias intra-cluster podem ser diminuídas, criando Clusters mais específicos. Por fim, deve-se reposicionar o centróide, esta nova posição deve ser a média da posição de todos os pontos do cluster **[6]**. Os dois últimos passos são repetidos, iterativamente, até que seja obtida a posição ideal de cada centróide.

2.3 KNN

O método dos k vizinhos mais próximos (kNN, do inglês k *nearest neighbors*) é um dos métodos de classificação automática mais simples e eficazes já propostos. O KNN foi proposto por Fukunaga e Narendra em 1975.

A ideia principal do KNN é determinar o rótulo de classificação de uma amostra baseado nas amostras vizinhas advindas de um conjunto de treinamento. A variável k representa a quantidade de vizinhos mais próximos que serão utilizados para averiguar de qual classe a nova amostra pertence **[7]**.

O cálculo mais utilizado para a definição da distância entre os elementos é a distância euclidiana, descrita na Fórmula 4 **[8]**.

$$D_E(\mathbf{p},\mathbf{q}) = \sqrt{(p_1 - q_1)^2 + \cdots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

É possível realizar o cálculo da distância através de outros algoritmos. Para entender melhor, podemos fazer uma análise sobre o pseudocódigo do algoritmo.

```
1  inicialização:
2      Preparar conjunto de dados de entrada e saída
3      Informar o valor de k;
4  para cada nova amostra faça
5      Calcular distância para todas as amostras
6      Determinar o conjunto das k's distâncias mais próximas
7      O rótulo com mais representantes no conjunto dos k's
8      vizinhos será o escolhido
9  fim para
10 retornar: conjunto de rótulos de classificação
```

2.4 LOGISTIC REGRESSION

Em 1944 Joseph Berkson, começou o desenvolvimento do que se tornaria o *Logistic Regression*.

A regressão logística (LR, do inglês *Logistic Regression*) é uma técnica estatística que tem como objetivo produzir, a partir de um conjunto de observações, um modelo que permita a predição de valores tomados por uma variável categórica, frequentemente binária, em função de uma ou mais variáveis independentes contínuas e/ou binárias **[9]**.

Para fazer o treinamento é utilizado uma base de dados, com variáveis independentes, e a partir disso, é usado uma função logarítmica para calcular a probabilidade de o evento ser uma combinação linear das variáveis independentes **[9]**.

A partir disso é possível prever a classificação de novas entidades no Data Space.

2.5 TRABALHOS RELACIONADOS

2.5.1 *Customer churn prediction system: a machine learning approach*

Este trabalho apresenta um sistema de predição do *churn* na indústria de telecomunicações. Através de algoritmos de *machine learning* e classificação, o algoritmo proposto consegue atingir uma acurácia de 84% na identificação do *churn* em usuários de empresas de telecom **[10]**.

2.5.2 *B2C E-Commerce Customer Churn Prediction Based on K-Means and SVM*

Este trabalho apresenta uma análise sobre a base de dados em empresas de e-commerce da China. Através da aplicação de algoritmos de *machine learning*, combinada com a utilização de segmentação k-means e SVM, o algoritmo faz uma síntese sobre os dados da empresa Alibaba Cloud Tianchi. A ideia do artigo é permitir a classificação e identificação do perfil do consumidor que "desiste" da empresa **[11]**.

2.5.3 *Customer Churn in Retail E-Commerce Business: Spatial and Machine Learning Approach*

O estudo tem por objetivo prever o *churn* de clientes de uma plataforma de e-commerce do Brasil. A abordagem consiste em três estágios que são combinados para criar o perfil do cliente que abandonará a plataforma. Foram combinados dados da empresa, informações sócio geográficas e textos de revisão. Foram aplicadas técnicas de regressão logística e gradientes de classificação para tentar prever o *churn* de usuários **[12]**.

3 MATERIAIS E MÉTODOS

3.1 DESCRIÇÃO DA BASE DE DADOS

Na Tabela 1 constam dados dos fatos transacionais, contendo 236.006 (duzentos e trinta e seis mil e seis) transações. É possível obter dados que caracterizam e especificam cada uma das transações como, por exemplo, data da transação, status, produtos utilizados, tipo da operação de venda, valores referentes a receitas e custos.

Tabela 1: Dados transacionais.

NOME	DESCRIÇÃO
transacao_id	ID da transação
produto	Produto da transação
tipo_venda	Modo de pagamento da venda
data_transacao	Data da venda
status_transacao	Status da venda
parcela	Quantidade de parcela
bandeira	Bandeira da venda
modo_ recebimento	Modo de recebimento do lojista
valor_transacao	Valor Bruto da transação
valor_preco_mdr	Valor calculado de desconto que o lojista paga no MDR
valor_preco_ antecipacao	Valor calculado de desconto que o lojista paga na Antecipação
liquido_lojista	Líquido do lojista daquela transação
cet_lojista	Custo Efetivo Total do lojista (percentual de desconto)
valor_custo_mdr	Valor calculado de desconto que a Justa paga no MDR
valor_custo_ antecipacao	Valor calculado de desconto que a Justa paga na Antecipação
receita_total_ mdr	Receita total da Justa no MDR
receita_total_ antecipacao	Receita total da Justa na Antecipação
receita_total	Receita Total (MDR + Antecipação)
lojista_id	ID do lojista

Fonte: Os autores.

A Tabela 2 contém dados sobre os clientes, através dela é possível obter identificação do cliente, definição se é pessoa física ou jurídica (CPF ou CNPJ), localização e elementos a nível transacional, como status, data de cadastro, primeira e última movimentação na empresa. A base conta com 3.402 (três mil quatrocentos e duas) entradas, sendo 2.973 (dois mil novecentos e setenta e três) entradas de clientes ativos.

Tabela 2: Dados dos clientes.

NOME	DESCRIÇÃO

DOI: 10.25286/repa.v7i3.2461

est_id	ID do lojista
est_status	Status do lojista
est_document_ type	Tipo documento
est_mcc_code	Código segmento profissional
est_mcc_name	Nome segmento profissional
estado	Estado do lojista
cidade	Cidade do lojista
data_cadastro	Data de cadastro do lojista
primeira_mov	Primeira movimentação do lojista
ultima_mov	Última movimentação do lojista
hunting	Canal de entrada do lojista
celula	Célula do canal de entrada do lojista

Fonte: Os autores.

A Tabela 3 apresenta a base de dados, após a realização da etapa de concatenação da tabela de clientes com a tabela transacional.

Tabela 3 Dicionário de Dados da Base Final.

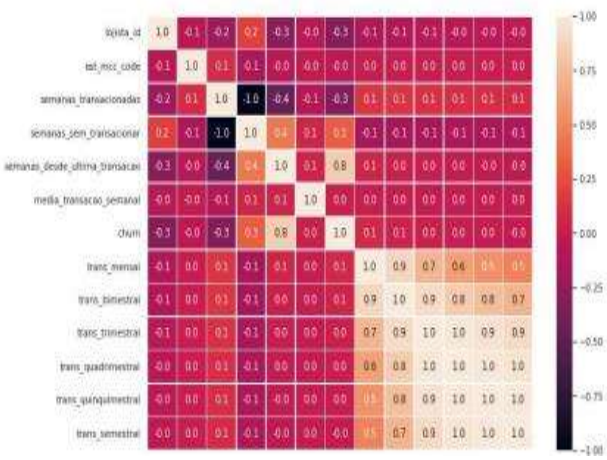
NOME	DESCRIÇÃO
est_id	ID do lojista
est_status	Status do lojista
est_mcc_code	Código segmento profissional
num_operacoes	Número de transações de um cliente
churn	se o usuário é churn ou não
semanas_ transacionadas	Quantidade de semanas com transações do usuário
semanas_sem_ transacionar	Quantidade de semanas sem transações do usuário
semanas_desde_ ultima_transacao	Quantidade de semanas desde a última transação do cliente
media_transacao_ semanal	Valor médio de das somas das transações durante a semana

Fonte: Os autores.

3.2 ANÁLISE DESCRITIVA DOS DADOS

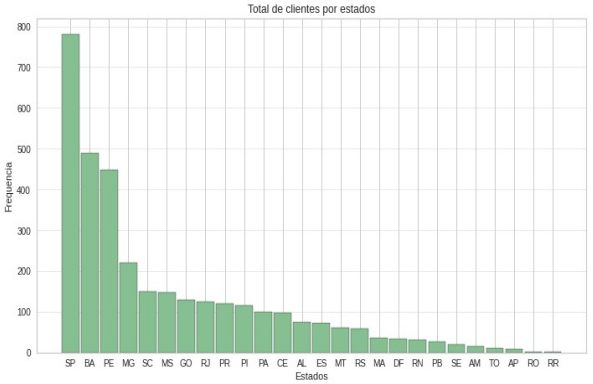
A partir da base de dados final foi possível criar gráficos que apresentam informações acerca dos dados. As figuras de 1 a 4 apresentam a análise descritiva dos dados.

Figura 1 - Matriz de correlação dos dados.



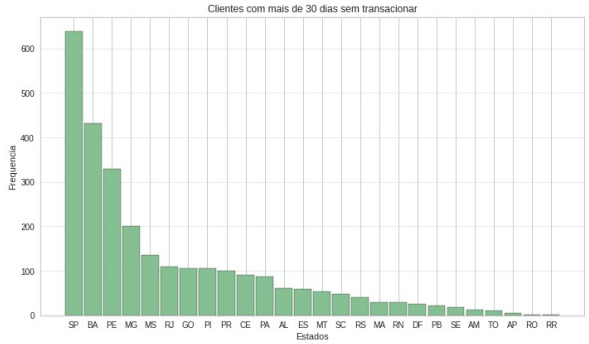
Fonte: Os autores.

Figura 2 - Distribuição de clientes por estado



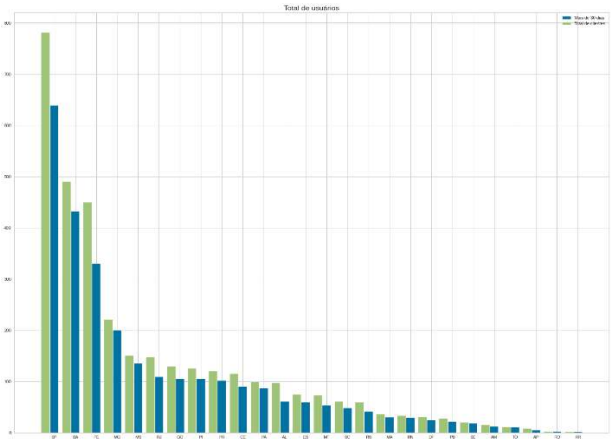
Fonte: Os autores.

Figura 3 - Clientes sem transacionar por mais de 30 dias



Fonte: Os autores.

Figura 4 - Relação entre número de clientes e interrupção de utilização



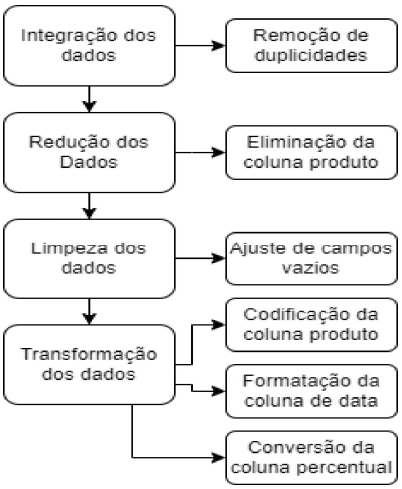
Fonte: Os autores.

3.1 PRÉ-PROCESSAMENTO DOS DADOS

O pré-processamento é um conjunto de tarefas que serão responsáveis por preparar, organizar e estruturar os dados. É fundamental para a realização de análises e predições.

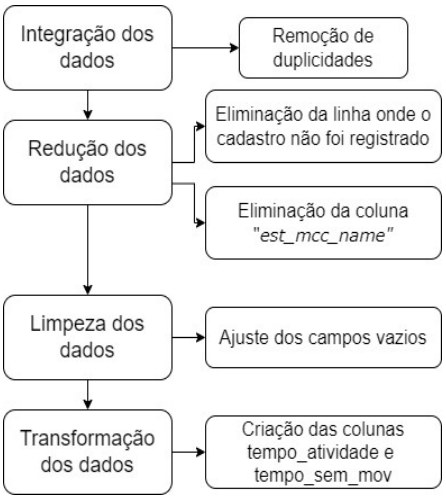
Dadas as bases sobre fatos transacionais e dimensão dos lojistas, as etapas de pré-processamento foram realizadas como mostram as figuras de 5 a 6.

Figura 5 - Fluxograma de pré-processamento da base de fatos transacionais.



Fonte: Os autores.

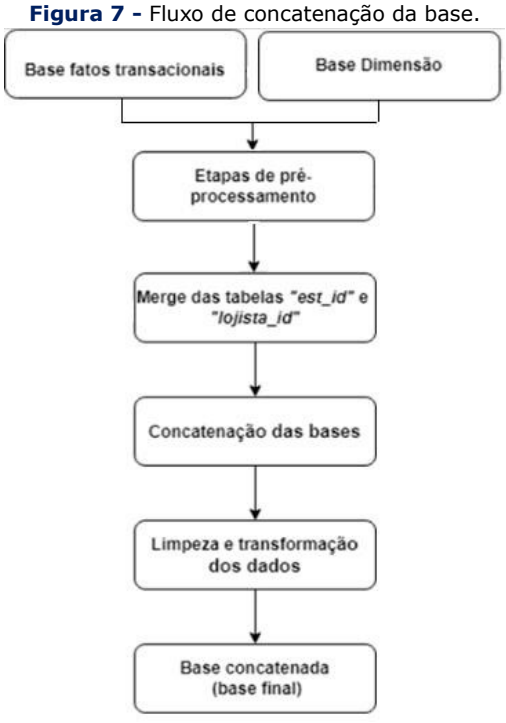
Figura 6 - Fluxograma de pré-processamento da base de dimensão dos lojistas.



Fonte: Os autores.

Para tornar o acesso mais eficiente, foi realizada a concatenação das bases, de modo que todos os dados relevantes para o processo foram unificados.

Figura 7 - Fluxo de concatenação da base.

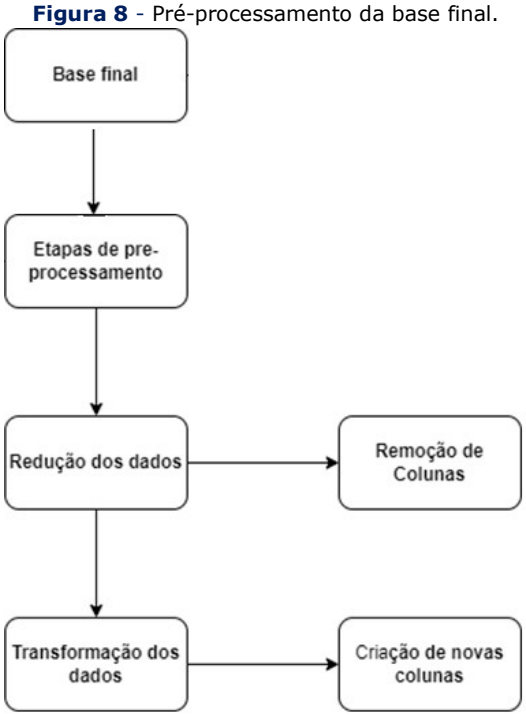


Fonte: Os autores.

Seguindo a proposta do CRISP-DM e o fluxo interativo de análises, foi necessário realizar novos pré-processamentos na base concatenada. Deste modo, o fluxo segue de acordo com a Figura 8.

DOI: 10.25286/repa.v7i3.2461

Figura 8 - Pré-processamento da base final.

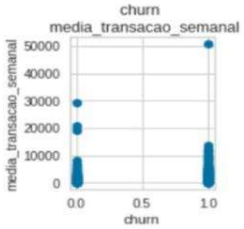


Fonte: Os autores.

4 RESULTADOS

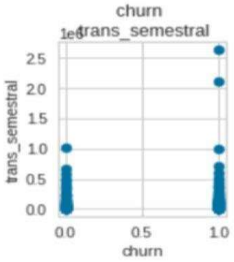
Para o cenário de análise da base de crédito e débito sem separação por tipo de usuário, os resultados podem ser analisados nas figuras de 9 a 18.

Figura 9 - Resultado do K-means, com a base total, para churn de acordo com a média das transações semanais.



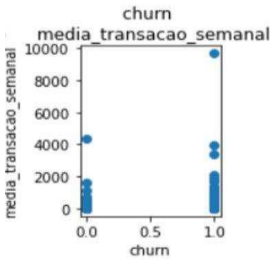
Fonte: Os autores.

Figura 10 - Resultado do K-means, com a base total, para churn de acordo com as transações semestrais.



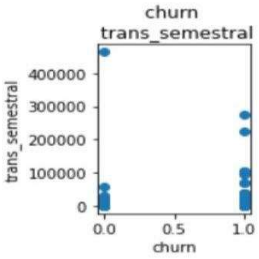
Fonte: Os autores.

Figura 11 - Resultado do K-means, com usuários de lojas têxtil, para churn de acordo com a média das transações semanais.



Fonte: Os autores.

Figura 12 - Resultado do K-means, com usuários de lojas têxtil, para churn de acordo com as transações semestrais.



Fonte: Os autores.

Figura 13 - Matriz de confusão do KNN para a base completa, acurácia de 74%.

188	85
171	578

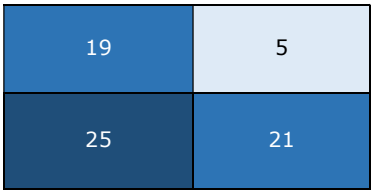
Fonte: Os autores.

Figura 14 - Matriz de confusão do KNN para usuários de lojas têxtil, acurácia de 81%.

17	9
6	49

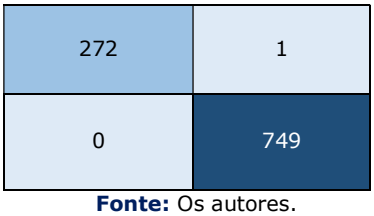
Fonte: Os autores.

Figura 15 - Matriz de confusão do KNN para usuários de lojas com finalidades educacionais, acurácia de 57%.



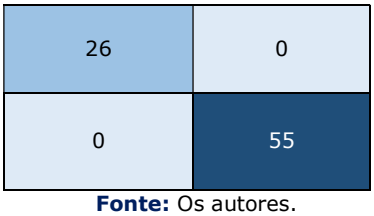
Fonte: Os autores.

Figura 16 - Matriz de confusão da Logistic Regression para a base completa, acurácia de 99%.



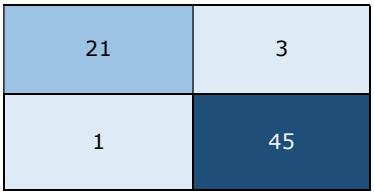
Fonte: Os autores.

Figura 17 - Matriz de confusão da Logistic Regression para a para usuários de lojas têxtil, acurácia de 100%.



Fonte: Os autores.

Figura 18 - Matriz de confusão da Logistic Regression para a lojas com finalidades educacionais, acurácia de 94%.



Fonte: Os autores.

Os agrupamentos gerados pelo K-means, mostram que não existe um padrão nas transações dos clientes, impossibilitando afirmar algo sobre os perfis de clientes que virão a ser um churn. Já os resultados obtidos nos classificadores, mostram que com as transações do cliente consegue-se informar se estes vão parar de transacionar ou não.

5 CONCLUSÕES

A falta de transações de clientes numa empresa é um problema muito frequente a ser lidado. Prever se um cliente é um possível churn ou não se mostra relevante para estabelecer técnicas que evitem que os clientes parem de realizar atividades transacionais na empresa [2].

Como solução do projeto, foi implementado um algoritmo de K-means para agrupar os dados estabelecidos e então dividir em dois grupos para serem analisados os resultados para os grupos formados. Também foram implementados dois algoritmos de classificação (KNN e Logistic Regression) estes responsáveis por indicar se, de acordo com os dados transacionais fornecidos, o cliente é ou não um churn.

Os resultados desse estudo possuem algumas limitações, pois foi utilizada uma base de dados real em que aproximadamente 81% dos clientes pararam de transacionar.

A fim de solucionar o problema, foram analisados dois cenários, os de transações de débito e crédito e as transações dos setores de loja têxtil e do setor educacional.

Os resultados obtidos nos agrupamentos, mostram que a base analisada não possui padrões suficientes para agrupar os dados em perfis de clientes que seriam necessários para a nossa análise, impossibilitando assim, observar os comportamentos de determinados tipos de clientes.

Os classificadores, por sua vez, conseguem solucionar a previsão dos resultados, para os cenários analisados. Na análise de crédito e débito, utilizando o KNN a acurácia é de 75%, sendo um resultado sólido e bom para classificar um cliente. Já nas análises dos setores de loja têxtil e educacional, obteve-se as acurácias de 81% e 57%, respectivamente, isso se deve à distribuição dos setores na base, setores de lojas têxteis são a maioria, enquanto educacional, não.

O LR, consegue uma acurácia de 99% para a análise de crédito e débito, enquanto para os casos de loja têxtil e educacional são, respectivamente, 100% e 94%.

De maneira geral, este estudo mostra que a base analisada possui muita rotatividade de clientes o que impossibilita uma análise precisa e clara do churn. Para o escopo estimado, o projeto consegue dizer se um cliente é churn, com base nas suas transações. Porém, para uma análise mais elaborada dos perfis de cliente e resultados

DOI: 10.25286/repa.v7i3.2461

melhores, se torna necessário uma base com maiores informações sobre o cliente em si, como renda mensal, ocupação entre outros.

REFERÊNCIAS

[1] KLEPAC et al. **Developing churn models using data mining techniques and social network analysis**, SCOPUS, 2015.

[2] GOLD C. S. **Fighting churn with data: The science and strategy of customer retention,** Manning Publications, 2020.

[3] OXFORD UNIVERSITY. **Oxford Dictionary of English**, 2013.

[4] MEHTA, N.; STEINMAN, D.; MURPHY, L. **Customer Success: How Innovative Companies Are Reducing Churn and Growing Recurring Revenue**, Editora Wiley, 2016.

[5] WU, J. **Advances in K-means Clustering: A Data Mining Thinking**, Springer, 2012.

[6] NUNES, D. **Um breve estudo sobre o algoritmo K-means**, Dissertação de Mestrado, Universidade de Coimbra, 2016.

[7] KRAMER, O. **Dimensionality Reduction with Unsupervised Nearest Neighbors**, Springer, 2013.

[8] TAN, P.; STEINBACH, M., KUMAR, V. **Introdução ao Data Mining - Mineração de Dados**, Ciência Moderna, 2009.

[9] MENARD, S.; **Logistic Regression: From Introductory to Advanced Concepts and Applications**, SAGE Publications, Inc., 2010.

[10] LALWANI, P. et al. **Customer churn prediction system: a machine learning approach**. Computing 104, 271–294 (2022).

[11] XIAHOU, X.; HARADA, Y. **B2C E-Commerce Customer Churn Prediction Based on K-Means and SVM**. Jurnal of Theorical. Applied Electronic Commerce Research, 2022, 17(2), 458-475.

[12] MATUSZELAŃSKI K.; KOPCZEWSKA K. **Customer Churn in Retail E-Commerce: Business: Spatial and Machine Learning Approach**. Journal of Theoretical and Applied Electronic Commerce Research, 2022, 17(1):165-198.