Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Tatiane Alves

Mineração de dados para redução efetiva de CAC

Curitiba
2019

Tatiane Alves

Mineração de dados para redução efetiva de CAC

Monografia apresentada ao Programa de Especialização em *Data Science* e *Big Data* da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Prof. André Ricardo Abed Grégio

Curitiba
2019

Mineração de dados para redução efetiva de CAC

Tatiane Alves[1]

[1]Aluna do programa de Especialização em Data Science & Big Data

Resumo

O CAC, custo de aquisição de clientes, é uma métrica utilizada por empresas para o acompanhamento financeiro de campanhas de marketing digital. O escopo do CAC vai da atração de um visitante a uma página até o fechamento de uma compra, podendo gerar inúmeros dados. Esses dados (e quanto mais completos, diversificados eles forem) permitem realizar o cálculo do CAC e tomar decisões efetivas de melhoria a partir dos resultados obtidos em uma campanha de marketing digital. O sucesso atrelado às ações adotadas pelas organizações baseia-se fundamentalmente da contribuição de cada cliente adquirido em relação ao valor total do mesmo. Uma vez que as campanhas envolvem investimentos e as empresas possuem metas periódicas, o cumprimento dessas e a maximização dos lucros envolve aumento nas vendas com a base de clientes vigente ou atração de novos clientes. Portanto, é preciso conhecer a fundo os clientes, o melhor meio de atraí-los e retê-los, além de se planejar campanhas com resultados efetivos. Para tanto, faz-se necessário um processo eficiente, inteligente e automatizado, o qual permita a análise dos clientes, segmentação por grupos e consequente auxílio na tomada de decisões para as próximas campanhas. Assim, neste trabalho propõe-se um sistema, cujo objetivo é auxiliar na redução do CAC via técnicas de aprendizado de máquina. A metodologia a ser utilizada é a RFM em conjunto com algoritmos de machine learning, em um estudo de caso com dados reais, segmentar os clientes por perfil de consumo e aplicar algoritmos que "aprendam"as preferências de cada grupo de clientes, de forma a se produzir insumos que permitam o lançamento de campanhas de marketing direcionado especificamente a esses clientes.

Palavras-chave: custo de aquisição de clientes, CAC, RFM

Abstract

The CAC, cost of acquiring customers, is a metric used by companies for the financial monitoring of digital marketing campaigns. The scope of the CAC goes from attracting a visitor to a page until the closing of a purchase, and can generate countless data. These data (and how much more complete, diversified they are) allow you to perform the CAC calculation and make effective improvement decisions from the results obtained in a digital marketing campaign. The success of the actions adopted by the organizations is fundamentally based on the contribution of each client acquired in relation to the total value of the same. Since campaigns involve investments and companies have periodic goals, meeting these and maximizing profits involves increasing sales with the existing customer base or attracting new customers. Therefore, it is necessary to know the customers in depth, the best way to attract and retain them, and to plan campaigns with effective results. For this, an efficient, intelligent and automated process is necessary, which allows the analysis of the clients, segmentation by groups and consequent aid in the decision making for the next campaigns. Thus, this work proposes a system, whose objective is to assist in the reduction of CAC through machine learning techniques. The methodology to be used is RFM in conjunction with machine learning algorithms, in a case study with real data, segment clients by consumption profile and apply algorithms that "learn" the preferences of each group of clients, in order to if it produces inputs that allow the launch of marketing campaigns directed specifically to these customers.

Keywords: customer acquisition, CAC, RFM

1. Introdução

O sucesso das ações de marketing adotadas pelas organizações baseia-se fundamentalmente na contribuição de cada cliente adquirido em relação ao valor e custo total do mesmo para a empresa [6]. É preciso conhecer a fundo os diferentes perfis de clientes do negócio, tanto quanto o melhor meio para atraí-los e retê-los, para se planejar campanhas direcionadas ao perfil em questão, obtendo consequentemente resultados mais

efetivos e com impacto financeiro positivo dentro da organização.

Com o advento da tecnologia e o aumento da quantidade dos canais de comunicações possíveis de serem utilizados para captação de potenciais clientes e retenção de clientes efetivos, tornou-se importante, como estratégia de diferencial competitivo organizacional, a adoção de posicionamento mais certeiro das empresas frente aos clientes ideias para o negócio.

Campanhas de comunicação são disparadas periodicamente pelas equipes de marketing, muitas vezes diária ou semanalmente, com objetivo de captar leads e aumentar o volume vendas do negócio. A construção dessas campanhas toma como base análises específicas e pontuais, muitas vezes realizadas por profissionais de marketing embasados tanto em dados quantitativos quanto qualitativos internos e externos ao negócio, visando a construção das potenciais personas a serem atingidas.

A subjetividade da análise realizada, em conjunto com as limitações da capacidade de processamento de grandes volumes de dados por um ser-humano [2], quando comparado ao poder de processamento de uma máquina, traz consigo limitações na identificação dos leads ideias do negócio junto ao mercado externo.

O CAC, ou custo de aquisição de clientes, é uma das métricas utilizadas pelas empresas brasileiras para o acompanhamento do desembolso financeiro em relação a investimentos em marketing digital. O escopo do CAC engloba a soma de gastos desde voltados à atração de visitantes, por meio de campanhas de comunicação em redes sociais por exemplo, até custos associados às ações realizadas ao longo do funil de vendas objetivando o fechamento de uma compra [6].

A contabilização desses gastos em conjunto com o número de clientes efetivos adquiridos, permite o cálculo do CAC e traz auxílio às tomadas de decisões visando melhoria das estratégias adotadas atualmente e resultados alcançados.

Para tanto, faz-se necessário um processo eficiente, inteligente e automatizado, que permita a análise dos clientes, por meio da segmentação dos grupos de personas com base em seu perfil de consumo, e auxilie no processo de tomada de decisões para as campanhas futuras.

Assim, neste trabalho propõe-se um sistema cujo objetivo é auxiliar na redução do CAC via técnicas de aprendizado de máquina. O projeto visa embasar a construção de uma aplicação web para visualização dos resultados, por meio da aplicação de técnicas de machine learning (ML) em um estudo de caso com dados reais de organização do setor de varejo brasileiro, que proponha um modelo possível de ser traduzido em plano de ações organizacionais para ajuda no alcance das metas organizacionais.

Para isso, escolheu-se a aplicação da técnica de marketing RFM em conjunto com o processo de data mining aqui proposto. RFM permite a construção dos segmentos de personas com base em seus perfis de consumo, para posterior personalização em relação de serviços e produtos ofertados pela empresa para cada um deles [1].

RFM é defendida pelos estudiosos como sendo um bom método de segmentação de clientes, que independe da quantidade de registros históricos utilizados durante a análise, trazendo informações úteis sobre clientes novos e atuais [1].

Por meio da análise dos dados históricos de compras dos consumidores, a metodologia propõe a utilização de três features para embasamento do seu resultado, sendo elas: *recência*, ou seja, o quão recentemente um consumidor realizou a última compra; *frequência*, como sendo quantas vezes, num determinado período de tempo, o mesmo comprou; e *valor*, como sendo o quanto ele comprou, em valores monetários, no mesmo período [1]. Clientes que compram mais recentemente, com maior frequência e valor, possuem maior score RFM e são consequentemente considerados os que mais trazem impacto financeiro para a organização.

Hoje em dia, RFM vem sendo aplicada em conjunto com técnicas de clusterização, classificação e associação vindas da ML, objetivando providenciar inteligência de mercado. Cases, como a utilização do K-means e RFM para a apoio na gestão de relacionamento com o cliente (CRM) e consolidação de fidelização dos clientes, podem ser encontradas frequentemente em estudos [1].

2. Processo de descoberta de conhecimento (KDD)

Como objetivo primário (Y) deste estudo a ser alcançado por meio da construção de uma solução inteligente propõe-se: reduzir custo de aquisição de clientes advindos de campanhas com marketing digital. No entanto, podemos considerar como labels alternativos (Ys): (i) aumento de 0,5% para 10-12%, no número de leads engajados; (ii) aumento do market share em paralelo a redução de custos com campanhas de marketing;

(iii) identificação das personas com mais representatividade no negócio, sendo esse último o foco do nosso trabalho.

Para isso, foi necessário um processo de extração de conhecimento que fosse independente do modelo aplicado para o alcance de nosso objetivo (Ys:iii). Logo, foi utilizado o processo de descoberta de conhecimento (KDD), para construção dos algoritmos inteligentes. Na Fig. 1. é possível ver a visão geral dos passos do processo KDD [2].

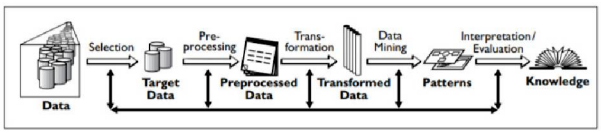

Figura 1: Fluxograma do processo de extração de conhecimento a partir de dados conhecido como KDD.

Também foi desenvolvida infraestrutura de TI para facilitar nossa pesquisa, que inclui uma base de dados, servidor cloud e API RESTFUL para tráfego dos dados. Foi utilizado um conjunto de medições para quantificar o desempenho das técnicas de mineração de dados aplicadas na redução do CAC, resultantes das diferentes ferramentas de modelagem aplicadas.

O presente trabalho propõe a utilização de algoritmo de clusterização, o K-means++, e de classificação, como árvore de decisão e KNN, em conjunto com a técnica RFM.

A decisão da utilização de modelagem de clusterização não supervisionada, K-means++, em detrimento das outras, se deu devido a vantagens em termos de tempo de execução e qualidade da clusterização em grandes datasets [1]. Logo, mesmo que o dataset utilizado neste trabalho seja *small data* e não *big data*, foi mantido a decisão da aplicação do K-means++ no caso de escalabilidade do dataset utilizado.

K-means++ em conjunto com RFM, objetiva encontrar segmentos de clientes com scores RFM similares de forma automatizada e em tempo real, sem a influência das regras de negócio utilizadas pela Empresa [1]. Isso permite encontrar uma solução algorítmica que não tenha a subjetividade da análise humana no processo de decisão.

Para fins de estudo de modelos baseados na segmentação atual adotada pela empresa, a mesma considera que os 5% primeiros clientes, com maior RFM score, são considerados melhores, sendo classificados como *'rockstar'*, o restante são segmentados em *'medium'* e *'low'*. Confrontamos essa metodologia com resultados obtidos pelo clusterizador e por três classificadores: árvore de decisão e KNN. Esses últimos foram treinados com base nos scores RFM levantados inicialmente. O objetivo foi encontrar o melhor classificador para os novos clientes com base em seu perfil de consumo, sem a necessidade de retreinamento do modelo.

2.1. Aquisição dos dados

Para que fosse possível o desenvolvimento do estudo proposto, foi utilizado para análise, dados de empresa brasileira do setor de varejo, no segmento Pet, aqui denominada Empresa, com dois anos de histórico de dados. Os dados selecionados para a construção da ideia proposta são advindos de base de dados relacional do sistema de gestão de empresas (ERP), adotado pela organização para controle, gerenciamento e planejamento de suas operações.

2.1.1. Fluxograma de coleta

O desenvolvimento de API RESTFUL foi necessário para aquisição dos dados junto ao ERP. Por meio de script C, para cada tabela desejada, a API executa classe de busca dos dados. No mesmo script, é realizado o insert de dado em banco SQL Server. A Fig. 2 ilustra o processo criado.

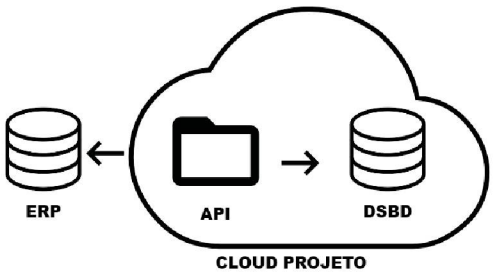

Figura 2: Processo de infraestrutura de servidor cloud, banco de dados SQL Server e API RESTFUL . Os dados são coletados da base de dados proposto do sistema ERP da empresa, e são enviados para a base de dados do presente projeto.

2.1.2. Detalhamento dos dados

Para esse estudo de caso, foi utilizado a base de dados proveniente dos módulos financeiro, em específico notas fiscais faturadas, e cadastrais, sendo cadastro de clientes e produtos. O dataset consiste de 6.858 registros de cadastro de clientes ativos e inativos e 1.799 registros de notas fiscais (NF) faturadas, sendo que os títulos das NF tanto quanto o detalhamento das mesmas estão em tabelas separadas e apresentam 2.031

registros de títulos e 16.966 registros de detalhamento de NF.

2.2. Limpeza e pré processamento dos dados

Foi realizado a construção de View, em nível de banco de dados, para disponibilização dos dados manipulados e pré-processados. No script da View, dados das tabelas bases cadastrais, como registros com duplicidade de CNPJ, status de cadastro cancelado, registros com nome de cliente vazio, foram eliminadas. É aplicado também a construção das features RFM. Fig. 3 esboça o fluxograma do processamento de dados aplicado.

Devido às características específicas de segmento de atuação da Empresa, da base resultante após manipulação, foram selecionados somente dados de clientes do tipo P.J. (pessoa jurídica). Reduzindo o dataset a ser trabalhado para 265 registros.

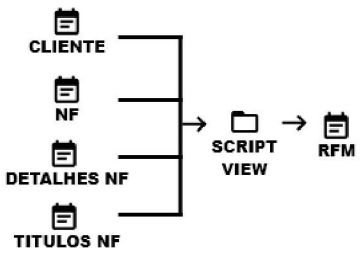

Figura 3: Fluxograma do processo de manipulação dos dados base por meio de Script SQL. As tabelas originais advindas do sistema ERP são manipuladas via script SQL e transformadas na view RFM.

2.2.1. Passo 1 - Reduzindo a dimensionalidade das features

A técnica RFM considera a construção de três features por cliente:

- Recência (R);
- Frequência (F);
- Valor (M);

Da base de clientes, dos 33 atributos vindos do ERP, analisamos somente o atributo *CodigoCliente*. Por meio de script SQL, foi realizado *join* de tabelas para trazer as NF e títulos por cliente. Das 21 features da tabela *TitulosNF*, foi utilizado somente um atributo, composto dos valor das parcelas das NF. A partir da soma desses valores obtém-se o valor em reais consumido pelo cliente. Para o cálculo da recência foi utilizado a coluna de data de emissão (*NfDEmi*) da última NF faturada em comparação com a data do momento da análise, e frequência foi formada pela totalização de NF emitidas no mesmo período.

Com objetivo de demonstrar a aplicação da análise RFM, a Tabela 1 mostra o primeiros 5 registros do dataset de dados de transação de clientes resultante da execução da View.

Tabela 1: Dataset resultante após redução de dimensionalidade de features.

Codigo Cliente	Valor (reais)	Frequência (número)	Recência (dias)
423521600	201175.96	7	61
423521616	8827.41	2	741
424949590	330.00	1	834
427042424	2042.84	8	394
431309320	472.63	2	435

2.2.2. Passo 2 - Ajustando para o objetivo alvo

Script python foi desenvolvido para consumir os dados vindos da *View* e processá-los com base na técnica RFM. Para cada cliente é atribuído score com base no seu histórico de comportamento junto a organização.

A técnica se inicia com a divisão do dataset estudado em 5 partes iguais, onde, para quintil, é atribuído score de 1 a 5. Após isso, para cada *feature RFM*, é realizado o processo de atribuição de score individualmente.

O dataset é ordenado primeiramente de forma crescente com base na coluna *Frequência*, onde para cada quintil atribuiu-se um score de 1 a 5, ficando os primeiros registros com score 5 e os últimos com score 1. É realizado o mesmo procedimento para as colunas *Valor* e *Recência*, sendo neste último caso realizado ordenação de forma decrescente [1].

É criado no dataset a coluna *RFMScore*, ou score RFM, por meio do cálculo da média dos scores atribuídos as colunas Recência, Frequência e Valor [**?**]. Com base na classificação dos perfis de clientes utilizados pela empresa atualmente, foi criado a coluna *Rotulo* no dataset com os labels estipulados pela Empresa.

A Tabela 1 abaixo mostra o resultado da aplicação do passo-a-passo da análise RFM.

Tabela 2: Resultado do dataset após aplicação da técnica RFM.

Codigo Cliente	R Score	F Score	M Score	RFM Score	Minha Classificacao
423521600	5	5	5	5	rockstar
423521616	5	4	5	4.6	rockstar
424949590	5	3	4	4	rockstar
427042424	5	2	4	3.6	rockstar
431309320	5	4	5	4.6	rockstar

2.3. Detalhes da modelagem

Para aplicação do *K-means*, foi escolhido o método *'k-means++'* default no *Scikit-Learn*. Foi mantido o valor *default* de 10 iterações com diferentes centroides do parâmetro *n_init*, sendo que o parâmetro *random_state* foi testado com 0 e 10, sendo mantido 10 para o presente trabalho.

A identificação do número ideal de clusters deste dataset se deu por meio da visualização via gráfico *Elbow*. A aplicação do gráfico *Silhoutte* permitiu o embasamento dessa escolha. Os *scores_silhoutte* resultantes tanto da aplicação do *K-means++* para 2 ou 3 clusters, ficaram muito abaixo de 1. Isso indica que os clusters encontrados em ambos os casos não estão muito bem separados, ou seja, estão relativamente próximos do limite de decisão dos seus vizinhos. Foi executado *K-means++* com parâmetro *n_clusters* igual a 3.

Para a aplicação dos classificadores, o dataset foi dividido em *set (X) & target (y)*, onde *X* consiste nas variáveis preditoras do dataset desconsiderando a coluna *Rotulo* e *y* a variável de saída representada pela coluna *Rotulo*. Para construção dos dados de teste e treinamento foi definido o tamanho do conjunto de testes com 30% de toda a base, e o restante dados de treinamento.

2.4. Resultados experimentais

O resultado da aplicação do *K-means++* foi comparado com o dado *Rotulo* como pode ser visto na Tabela 3 abaixo.

É possível vermos que clientes de perfis heterogêneos foram agrupados em um mesmo cluster, como visto no *cluster 1*. Nenhum cluster foi atribuído aos clientes *rockstar* ficando os mesmos totalmente misturados com clientes do tipo *medium* e *low*.

Tabela 3: Resumo da quantidade clientes por categoria em cada cluster retornado pelo K-means++.

Cluster	rockstar	medium	low
0	0	5	75
1	14	68	14
2	0	37	54

Matriz de confusão resultante da aplicação dos classificadores propostos podem ser vistas nas imagens abaixo. A Fig.4 mostra o resultado do aplicação do modelo Árvore de Decisão, executada pelo método *DecisionTreeClassifier* do *Scikit-Learn*. O mesmo conseguiu acertar suas classificações em sua totalidade.

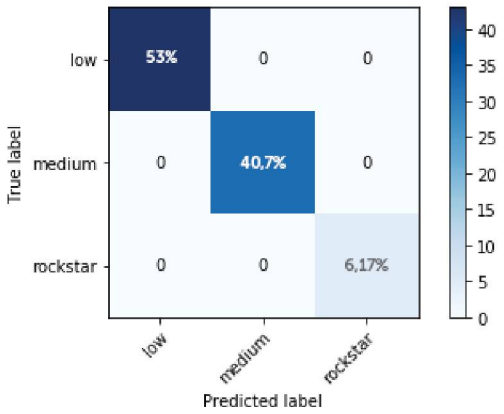

Figura 4: Matriz de confusão resultante do modelo Árvore de Decisão.

A acurácia do modelo Árvore de Decisão ficou em 100% e KNN teve como resultado 62%. Logo, o modelo de árvore de decisão foi o que apresentou melhores resultados em relação ao dataset utilizado. A precisão resultante no melhor caso ficou de 1.0, sendo no pior caso 0.44. A Fig.5 mostra a matriz de confusão resultante do modelo KNN.

3. Conclusão

A aplicação de técnicas de marketing defendidas pela literatura em associação a modelos inteligentes advindos da inteligência artificial nos mostra que existe uma grande dependência com outras variáveis para o alcance dos resultados esperados, seja a qualidade dos dados disponíveis, seja o segmento de atuação da empresa estudada.

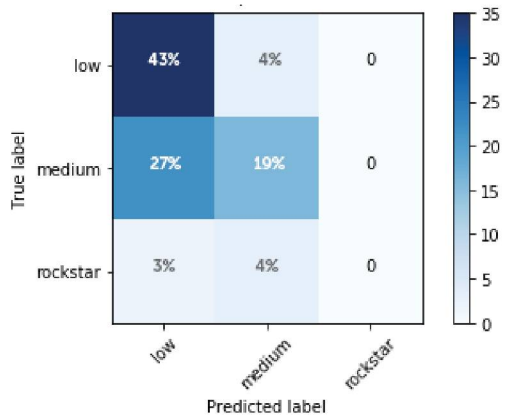

Figura 5: Matriz de confusão resultante do modelo KNN.

A metodologia RFM é encontrada na literatura com inúmeras variações em seu modelo de aplicação, podendo mudar as regras de aplicação de caso para caso. Sofrendo variações em relação ao perfil da organização a qual está sendo aplicada, por meio da necessidade de atribuição de pesos para os scores atribuídos aos atributos recência, frequência e valor. Isso torna-se um desafio quando se fala na construção de aplicação web de auxílio à tomada de decisão das equipes de marketing das empresas brasileiras.

Em relação aos resultados alcançados com a aplicação dos modelos de machine learning, o resultado da aplicação do cluster pode ser possivelmente justificado pelo volume de dados históricos disponíveis para o modelo. A aplicação também da técnica RFM em empresas do exterior, realizada pelos papers acadêmicos utilizados para embasamento do presente projeto, pode ser também um dos fatores influenciadores quando replicado a mesma metodologia para o cenário interno brasileiro. A adoção do classificador árvore de decisão se mostra pertinente para o objetivo proposto no presente trabalho.

Como proposta de evolução do presente trabalho, indica-se o estudo mais a fundo e comparativo entre as técnicas de aprendizado de marketing de clusterização que melhor atendam o perfil dos dados de empresas característica do varejo no brasil, respeitando as particularidades do negócio e necessidade de adaptação da aplicação do RFM. Traz também a oportunidade da construção de novos perfis de consumo possivelmente não identificados pela Empresa.

Agradecimentos

Gostaria de agradecer ao meu orientador, Dr.André Grégio, por todo o apoio e suporte para conclusão e superação dos desafios do presente trabalho projeto. Minha família, obrigada pela paciência e motivação. Ao meu sócio, Paulo Gemniczak, pela compreensão e suporte ao longo deste ano.

Referências

[1] D. Birant, *Data Mining Using RFM Analysis*, (Dokuz Eylul University, Turkey, 2011, p. 91-108), Knowledge-Oriented Applications in Data Mining, Prof. Kimito Funatsu (Ed.), InTech.

[2] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, *The KDD Process for Extracting Useful Knowledge from Volumes of Data*, Communication of the acm, November 1996, Vol. 39, nº. 1.

[3] *Apply RFM principles to cluster customers with K-Means*, Towards Data Science, https://towardsdatascience.com/apply-rfm-principles-to-cluster-customers-with-k-means-fef9bcc9ab16, Vallantin, W. L..

[4] D. D. Nimbalkar, P. Sha, *Data mining using RFM Analysis*, International Journal of Scientific Engineering Research,December 2013, Vol. 4, Issue 12.

[5] *Segmentação com a análise RFM*, Data Science com Marketing, https://medium.com/data-science-com-marketing/segmenta%C3%A7%C3%A3o-com-an%C3%A1lise-rfm-623100279269, Picoloto, L.

[6] *Custo de Aquisição de Clientes: entenda o que é e como reduzir o CAC da sua empresa*, RockContent, https://rockcontent.com/blog/custo-de-aquisicao-de-clientes/, Mesquita. R.

[7] A. Zheng and A. Casari, *Feature Engineering for Machine Learning*, United State of America, 2018, 1 ed..