Aplicação de Mineração de Dados para Detecção de Potenciais Churns em Empresa do Segmento SAAS

Leonardo Lucas de Melo
Laboratório de Inteligência Aplicada
Universidade do Vale do Itajaí
Itajaí, SC Brasil
leolucasm@edu.univali.br

Rafael Ballottin Martins
Laboratório de Inteligência Aplicada
Universidade do Vale do Itajaí
Itajaí SC Brasil
ballottin@univali.br

ABSTRACT

This paper tackles the problem of companies that offer subscription services, plans or any other recurring method of payment. In this commercialization model, it is important to keep the churn rate low, but to define strategies for reduce churn rate it is needed to identify what are the main reasons for losing customers. The application of data mining techniques may assist to find out patterns that can trace the most likely customers to become churns. Through this research, using the data mining process, it was possible to identify that, among other factors, the non-utilization of the main system modules and the high default rates corroborates for customers to become churn.

KEYWORDS

KDD, Churn, Data Mining

1 Introdução

O avanço tecnológico tem permitido a criação de diversas bases de dados com diferentes origens, desde comercial até científica. Essas bases de dados podem fornecer conhecimentos relacionados ao perfil de clientes e do negócio. Porém, nem sempre essas informações são fáceis de serem identificadas, analisar um conjunto de dados apenas por meio de recursos humanos pode ser complexo [1].

O *churn* é um termo utilizado para descrever o cliente que encerra o contrato com uma empresa para consumir os produtos ou serviços dos concorrentes. Se a empresa quiser impedi-lo de partir, terá que realizar alguma ação preventiva ou de retenção. Para isso, é necessário entender os motivos que levaram aquele cliente a tomar tal decisão [2].

O desenvolvimento deste projeto teve como objetivo traçar o perfil de clientes que potencialmente se tornarão *churn*, por meio da utilização do processo de *Knowledge Discovery in Databases* (KDD) [3]. A pesquisa foi feita com os dados de uma empresa do segmento SAAS, esse segmento é o principal foco deste trabalho.

3 Solução Proposta

Nesta pesquisa, foi aplicado o processos de KDD na base de dados da empresa HiGestor, que atua no segmento de desenvolvimento e comercialização de software SAAS.

As técnicas de mineração de dados foram aplicadas sobre os dados relacionados ao principal software da empresa. Este software tem o propósito de facilitar e simplificar os processos de gestão de instituições como sindicatos, associações e federações. Por meio do sistema é possível realizar a emissão de boletos, realizar a gestão de contribuintes e eventos. Além disso, também permite o controle e acompanhamento do balanço financeiro das instituições.

Objetivou-se, por meio do processo de KDD, descobrir padrões e traçar o perfil dos clientes da empresa HiGestor que potencialmente se tornarão churn.

4 Projeto

Para obter um melhor entendimento do negócio, foi realizada uma reunião com especialistas dos setores administrativo, comercial e de produto. Na sequência, iniciou-se o processo de extração dos dados disponibilizados pela empresa.

4.1 Conjunto de Dados

Os dados utilizados no KDD foram extraídos de três diferentes bases de dados. A primeira delas era oriunda de um software desenvolvido pela própria HiGestor, que contém todo o histórico dos seus clientes. A segunda base de dados continha informações referentes a todos os atendimentos prestados aos clientes. Por fim, a terceira base de dados era oriunda do software SaaS e continha informações referentes aos clientes e à utilização do sistema. Este sistema é utilizado principalmente para emissão de boletos e gestão financeira das entidades.

4.2 Análise dos Dados e Seleção de Ferramentas

Com o objetivo de obter um melhor conhecimento sobre as bases de dados disponíveis, foi realizada uma análise exploratória dos dados. Observaram-se as variáveis contidas, a distribuição dos dados e a qualidade das informações. Após a etapa de análise

exploratória, construiu-se um *Data Warehouse* (DW), integrando os 3 conjuntos de dados disponibilizados pela empresa.

A partir da realização de pesquisas, buscou-se identificar qual ferramenta poderia ser utilizada no processo de mineração de dados e optou-se por utilizar a ferramenta *Rapidminer*. Esta ferramenta contempla diversos algoritmos, dentre eles algoritmos para mineração de dados, mineração de texto e aprendizado de máquina.

4.3 Seleção e Transformação de Atributos

Após a análise exploratória do conjunto de dados, realizou-se uma reunião com os gestores da empresa. Nesta reunião foram definidos quais atributos do conjunto de dados poderiam ter influência sobre o *churn*. Foram selecionados 24 atributos que indicam padrões de utilização do sistema e perfil do cliente.

Na sequência, iniciou-se o processo de transformação. Alguns dos atributos extraídos precisavam de transformação para utilização nas tarefas de classificação e associação.

Dentre as transformações realizadas, a principal transformação foi sobre o atributo qtd_recebimentos_cadastrados. Ao realizar a etapa de exploração de dados foi identificado que o atributo qtd_recebimentos_cadastros, quando analisado de forma individual, não agrega nenhum tipo de conhecimento. Para definir se uma quantidade de recebimentos é baixa ou alta, por exemplo, é necessário avaliar a quantidade de filiados e associados que a entidade possuí. Para resolver este problema foi criado o atributo perc_recebimentos_cadastrados, que recebe o percentual de recebimentos cadastrados pelo cliente da empresa.

4.4 Seleção de Técnicas de Mineração de Dados

Para obter regras que apontassem os perfis de clientes que potencialmente se tornarão *churn*, optou-se por utilizar a técnica de classificação. Esta técnica, por meio da análise dos atributos relacionados ao perfil do cliente, pode criar classes que indicam os clientes que potencialmente cancelarão o contrato. Para esta pesquisa, o atributo utilizado como classe foi o *churn* [4].

Definiu-se que seria utilizado algum algoritmo de classificação que exibisse os resultados em forma de uma árvore de decisão. Esta escolha foi feita porque as árvores de decisões geram regras de classificações que são mais claras e fáceis de interpretar. Portanto, optou-se por testar os algoritmos *J48* e *Decision Tree*, porque estes algoritmos podem trabalhar tanto com atributos categóricos quanto numéricos [5].

Para identificar características em comum dos clientes que se tornaram *churn* e ponderar as regras de classificação, definiu-se que seria utilizada a técnica de associação. A tarefa de associação identifica os atributos que se correlacionam na base de dados, gerando regras que indicam o atributo consequente quando tais associações ocorrem. Após analisar trabalhos similares e realizar testes preliminares, optou-se por utilizar o algoritmo *Apriori* [6][7][8][9]. Para esta pesquisa, foram observadas principalmente as regras cujo atributo consequente foi o *churn*.

4.5 Aplicação dos Algoritmos de Mineração de Dados

Para a aplicação da tarefa de classificação dividiu-se o conjunto de dados em dois subconjuntos. O primeiro subconjunto com 70% dos registros foi utilizado para treinamento. O segundo subconjunto com os outros 30% foi utilizado para validar o modelo gerado.

Foram aplicados os algoritmos J48 e Decision Tree ao conjunto de dados, com o objetivo de descobrir quais deles poderiam gerar modelos mais precisos. Para realização dos testes com os algoritmos, foram definidos alguns parâmetros de configuração. Foram eles: Ganho mínimo, Tamanho mínimo da folha e Profundidade máxima.

Testes preliminares foram realizados e identificou-se que, para esta pesquisa, o ganho mínimo com índice de 0.30 gerou melhores resultados. Definiu-se que as folhas deveriam ter ao mínimo 2 ramificações, ou seja, para cada nó final da árvore deveriam existir ao menos duas saídas. Por fim, foi definido que a profundidade máxima da árvore deveria ser de 30 nós, evitando que fossem geradas árvores muito extensas.

Os resultados do algoritmo J48 apresentaram um percentual de acurácia de 96,58%, maior que os resultados do algoritmo Decision Tree, que apresentaram uma acurácia de 94,13%. Sendo assim, optou-se por utilizar o algoritmo J48 para geração da árvore de decisão.

A árvore de decisão gerou 27 regras de classificação, as principais regras estão listadas a seguir:

1. if limite_associados > 2000 THEN not churn
2. if limite_associados <= 2000 and usa_receita_facil = true THEN not churn
3. if limite_associados <= 2000 and usa_receita_facil = false and perc_recebimentos_pagos = MUITO_BAIXO and valor_contrato <= 160 and teve_acompanhamento_inicial = NÃO THEN churn

13. if limite_associados <= 2000 and usa_receita_facil = false and perc_recebimentos_pagos = MUITO_BAIXO and valor_contrato > 160 and teve_acompanhamento_inicial = SIM and qtd_tickets <= 20 and qtd_filiados > 10 THEN churn

19. if limite_associados <= 2000 and usa_receita_facil = false and perc_recebimentos_pagos = MEDIO and aviso_cobranca_a_vencer = true THEN not churn

21. if limite_associados <= 2000 and usa_receita_facil = false and perc_recebimentos_pagos = MEDIO and aviso_cobranca_a_vencer = false and

teve_acompanhamento_inicial = NÃO and limite_associados > 600 THEN churn

Para identificar variáveis do conjunto de dados que se correlacionavam na ocorrência de *churn*, foi utilizado o algoritmo *Apriori* da tarefa de associação. Foi necessário definir valores para dois parâmetros importantes, que são a confiança e o suporte mínimo.

A confiança foi estipulada como 75%, levando em considerações trabalhos similares e o conjunto de dados [6][7][8]. Para definir o suporte também foi necessário avaliar o conjunto de dados. Levando em consideração que 12,73% dos registros são referentes aos clientes que se tornaram *churn*, foi estabelecido um suporte mínimo de 0.0125, que representa 10% dos cancelamentos.

Para esta pesquisa, consideraram-se principalmente as regras cujo atributo consequente era o *churn*. A seguir estão listadas as principais regras:

3. limite_associados=[0, 2025], plano=Prata, acesso_portal=false, usa_eventos=false THEN churn = true (Suporte: 0.0137, Confiança: 0.95)
4. limite_associados=[0, 2025], plano=Prata, acesso_portal=false, perc_recebimentos_cadastrados =BAIXO THEN churn = true (Suporte: 0.0137, Confiança: 0.95)
9. limite_associados=[0, 2025], acesso_portal=false THEN churn = true (Suporte: 0.0297, Confiança: 0.95)
10. acesso_portal=false, perc_recebimentos_cadastrados=BAIXO THEN churn = true (Suporte: 0.0228, Confiança: 0.95)
11. limite_associados=[0, 2025], acesso_portal=false, usa_eventos=false THEN churn = true (Suporte: 0.0228, Confiança: 0.95)
12. acesso_portal=false, usa_eventos=false, perc_recebimentos_cadastrados =BAIXO THEN churn = true (Suporte: 0.0228, Confiança: 0.95)
13. limite_associados=[0, 2025], acesso_portal=false, usa_eventos=false, perc_recebimentos_cadastrados=BAIXO THEN churn = true (Suporte: 0.0228, Confiança: 0.95)

5 Conclusão

A ferramenta *Rapidminer* forneceu todos os recursos necessários nas etapas de pré-processamento e mineração de dados. Reuniões com os especialistas dos setores comercial, administrativo e de produto forneceram um melhor entendimento do negócio e dos problemas de pesquisa.

Para identificar os perfis de clientes que potencialmente se tornarão *churn*, aplicaram-se as tarefas de classificação e associação.

Utilizou-se o algoritmo *J48*, da tarefa de classificação, para geração de regras em formato de árvore de decisões. Os resultados do algoritmo J48 apresentaram uma acurácia de 96,58% e indicaram os principais motivos que levam a decisão de cancelamento de contrato. A partir da árvore de decisão gerada, descobriu-se que clientes mais estruturados, que possuem um porte maior, tem tendência a não se tornarem *churn*s. Além disso, foi percebido que a alta taxa de inadimplência colabora para a decisão de abandono do sistema.

Identificou-se que a não utilização de 4 módulos do software corrobora para que ocorra o *churn*. Os módulos que são exibidos nas regras de classificação são: Receita Fácil, Eventos, Aviso de Cobrança a Vencer e Aviso de Cobrança em Atraso. Todos estes módulos interferem diretamente no balanço financeiro dos clientes, ajudam a reduzir inadimplência, reduzir custos ou a gerar receitas extras.

Para descobrir quais atributos se correlacionavam nas ocorrências de *churn*, foi aplicado o algoritmo *Apriori*, da tarefa de associação. Foi estipulado que para a extração dos resultados, as regras de associação deveriam ter uma confiança mínima de 75% e suporte mínimo de 0.0125, que representa 10% dos cancelamentos.

Os resultados demonstraram que a não utilização de 2 módulos do sistema está associada a ocorrências de *churn*. Os módulos que aparecerem nas regras de associação são: Eventos e Portal do Associado. Os gestores do negócio entenderam que os clientes que não utilizam os principais módulos do sistema não conseguem maximizar os benefícios que o software pode oferecer.

Por fim, as descobertas deste trabalho permitem que a empresa inicie algumas ações para redução do *churn*. É possível definir estratégias de marketing buscando aumentar o engajamento dos clientes com os principais módulos do sistema. O marketing também pode gerar conteúdos que ajudem os clientes a encontrarem maneiras de reduzir o percentual de inadimplência. Além disso, a empresa pode planejar ações para dar um maior destaque aos principais módulos do sistema, demonstrando quais benefícios estes módulos podem gerar para os clientes que os utilizam.

REFERÊNCIAS

[1] Ronaldo Goldschmidt and Emmanuel Passos. DataMining - Um Guia Prático.v. 1, 2005.
[2] Nicolas Glady and Bart Baesens and Christophe Croux. Modeling churn using customer lifetime value. In: European Journal of Operational Research, v. 197, n. 1, p. 402–411, 2009.
[3] Daniel Larose. Discovering Knowledge in Data: An Introduction to Data Mining, 2005.
[4] Jiawei Han and Micheline Kamber and Jian Pei. Data Mining: concepts and techniques, 2011.
[5] Cassio Lorenzett and Alex Telocken. Estudo Comparativo entre os algoritmos de Mineração de Dados Random Forest e J48 na tomada de Decisão. In: Simpósio de Pesquisa e Desenvolvimento em Computação, v. 2, 2016.

[6] Antônio R. F. Junior. Aplicação de mineração de dados no gerenciamento do churn em startup do segmento saas. 2017. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação). Universidade do Vale do Itajaí, Itajaí, 2017.

[7] Pedro H. G. Nobre. Aplicação de Modelos de Data Mining para Previsão de Churn em Telecomunicações. 2016. Trabalho de Conclusão de Curso (Especialização) - PUC-RIO, Rio de Janeiro, 2016.

[8] Bruna C. P. Brum. Estimação da taxa de churn para clientes de uma seguradora baseado em técnicas de reconhecimento de padrões. 2016. Trabalho de Conclusão de Curso (Especialização). Faculdade de Engenharia Elétrica, PUC-RIO, Rio de Janeiro, 2016.

[9] Lívia Vasconcelos and Cedric Carvalho. Aplicação de Regras de Associação para Mineração. Technical Report - RT-INF_004-04 – Relatório Técnico de Novembro, 2004.