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

Identify what are the main reasons for losing customers is essential for companies that offer subscription services, plans or any other recurring method of payment. 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 naturezas, desde comercial até científica. Porém, frequentemente elas são utilizadas apenas em atividades triviais e como fontes de consultas [1].

Diante desse cenário, torna-se evidente que muitas empresas não sabem como utilizar seus dados para obter conhecimentos que possam auxiliar nas tomadas de decisões. Analisar o conjunto de dados apenas por meio de recursos humanos não é eficiente e dificilmente trará resultados satisfatórios.

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 consumidor a tomar tal decisão [2].

Utilizando o processo de *Knowledge Discovery in Databases* (KDD), buscou-se gerar novos conhecimentos, potencialmente úteis, em relação aos principais motivos que levam os clientes ao *churn*.

2 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 dos processos de KDD, descobrir padrões e traçar o perfil dos clientes da empresa HiGestor que potencialmente se tornarão *churn*.

3 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 conjuntos que dados disponibilizados pela empresa.

3.1 Extração dos Dados

Os dados utilizados no KDD foram extraídos de três bases de dados, que são manipuladas por diferentes sistemas. Uma base de dados é oriunda de um software desenvolvido pela própria empresa, chamado de IBFTask. Este software contém todo o histórico dos clientes com a HiGestor, como informações cadastrais e dados referentes aos pagamentos das mensalidades.
A segunda base de dados é alimentada pelo software Mysuite e contém informações referentes aos atendimentos prestados aos clientes. A terceira base de dados é oriunda do software HiGestor e contém informações referentes aos clientes e à utilização do sistema.
Após realizar a análise de cada conjunto de dados disponível, foi necessário executar o processo de extração das bases de dados e a integração em um DW. As etapas de integração e construção do

DW consideraram todas as informações relevantes para a identificação dos motivos que levam os clientes ao *churn*.
Com o auxílio dos especialistas dos setores administrativo e financeiro foi iniciado o processo de extração da base de dados. Os especialistas envolvidos nas etapas de consulta e extração possuem conhecimento sobre os processos dos demais setores, como suporte, *costumer success* e desenvolvimento.
Por fim, foi realizada a limpeza dos conjuntos de dados, buscando garantir a integridade dos valores armazenados.

3.2 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. Os principais critérios utilizados para pesquisa e definição da ferramenta foram:

1. Possuir licença de uso livre para projetos acadêmicos;
2. Disponibilizar uma plataforma web que permitisse a execução de algoritmos de mineração de dados. Esse item poderia ser útil caso os equipamentos utilizados nessa pesquisa não tivessem recursos suficientes para a execução de algum algoritmo;
3. Disponibilizar documentação;
4. Fornecer recursos que auxiliam a etapa de pré-processamento.

Considerando os critérios definidos optou-se por utilizar a ferramenta *Rapidminer*. O *Rapidminer* é uma ferramenta que contempla diversos algoritmos, dentre eles algoritmos para mineração de dados, mineração de texto e aprendizado de máquina.

3.3 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 resultado da Equação 1:

$$\frac{qtd\_recebimentos\_cadastrados}{qtd\_filiados + qtd\_associados} * 100 \quad (1)$$

3.4 Seleção de Técnicas de Mineração de Dados

Para atender o objetivo da pesquisa, definiu-se que seriam aplicadas tarefas de classificação e associação. A tarefa de classificação, por meio da análise dos atributos relacionados ao perfil do cliente, pode criar classes que indicam os clientes que potencialmente cancelarão o contrato.

Foi definido 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 fáceis de interpretar.
Optou-se por utilizar o algoritmo J48 para a aplicação da tarefa de classificação, porque este algoritmo pode trabalhar tanto com atributos categóricos quanto numéricos. Além disso, o algoritmo J48 é uma evolução dos algoritmos ID3, C4.5 e C5.0 [3].
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*.

3.5 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.

Após aplicação do algoritmo de classificação, foram obtidos alguns resultados preliminares e que indicam possíveis motivos para cancelamento de contrato. A seguir são listadas as principais regras:

1. 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;
2. if limite_associados <= 2000 and usa_receita_facil = false and perc_recebimentos_pagos = ALTO and faturamento_dia_base = ANUAL and usa_eventos = true then not churn;
3. if limite_associados <= 2000 and usa_receita_facil = false and perc_recebimentos_pagos = ALTO and faturamento_dia_base = ANUAL and usa_eventos = false then churn;
4. if limite_associados <= 2000 and usa_receita_facil = true then not 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. 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.

Após aplicação do algoritmo de associação, foram obtidos alguns resultados preliminares e que indicam variáveis associadas em ocorrências de *churn*. A seguir são listadas as principais regras:

1. limite_associados <= 2025, plano = Prata, acesso_portal = false, usa_eventos = false then churn = true (Suporte: 0.0137, Confiança: 0.95);
2. limite_associados <= 2025, plano = Prata, acesso_portal = false, perc_recebimentos_cadastrados = BAIXO then churn = true (Suporte: 0.0137, Confiança: 0.95);
3. acesso_portal = false, usa_eventos = false, perc_recebimentos_cadastrados = false then churn = true (Suporte: 0.0228, Confiança: 0.78);
4. limite_associados <= 2025, acesso_portal = false, perc_recebimentos_cadastrados = BAIXO then churn = true (Suporte: 0.0228, Confiança: 0.78);

Observando as regras preliminares, foram identificados alguns indícios que podem corroborar para que os clientes cancelem o contrato com a empresa.

Conforme pode ser visto nas regras, o limite de associados que os clientes possuem influencia a ocorrência de *churn*. Além disso, a não utilização de pelo menos dois módulos do sistema e o percentual baixo de recebimentos pagos interferem para a decisão de cancelamento de contrato

4 Considerações Finais

Aplicar o processo de KDD, para a extração de conhecimentos, foi imprescindível para o andamento desta pesquisa.

Foram realizados testes preliminares com os algoritmos de classificação e associação. A partir destes testes, foram identificados atributos que precisariam de transformações. Com base nos testes, também foram definidas as configurações a serem utilizadas nas tarefas de classificação e associação.

Para a aplicação do algoritmo de associação, foi definida uma confiança mínima de 75% e um suporte mínimo de 0.0125, que representa 10% dos cancelamentos. Para o algoritmo de classificação foram definidas configurações para extração dos resultados. Foi definido que o ganho mínimo de informação deveria ser de 0.30, o tamanho mínimo da folha deveria ser 2 e o critério de escolha foi definido como acurácia.

Resultados preliminares apontam principalmente que o alto percentual de inadimplência e a não utilização de dois módulos do sistema corroboram para o *churn*. Diante desses resultados, surgiu a hipótese de que a não utilização de outros módulos do sistema também pode influenciar para que os clientes cancelem o contrato com a empresa.

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 tem a opção de demonstrar já nas etapas de treinamento os principais módulos do sistema, bem como os benefícios que eles podem entregar aos clientes.

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] 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.