Previsão de Churn em Plataformas com Efeitos de Rede baseado em Árvores de Decisão

Fernanda Coelho de Queiroz

Faculdade de Engenharia Elétrica e Computação - UNICAMP

Key-words: Churn Prediction, Decision Tree, Non-Contractual, Two-Sided Market, Network Effects, Time-Series

fernanda.cdqueiroz@gmail.com

Introdução

Churn de clientes é o nome do evento em que o cliente termina seu relacionamento com uma empresa, impactando negativamente a geração de receita. Sua redução pode representar uma vantagem relevante, principalmente no caso de negócios multi-laterais com efeitos de rede, que distinguem-se por fornecer serviços para ao menos dois grupos distintos de usuários e cuja proposta de valor está atrelada ao tamanho e qualidade da suas diferentes bases de usuários. Este trabalho aplica árvores de decisão para endereçar a evasão de clientes em organizações com modelos de negócios não contratuais. O modelo desenvolvido torna possível a identificação precoce dos clientes mais propensos à abandonar o serviço, permitindo e direcionando ações preventivas.

Metodologia

Devido à variação temporal da taxa base de churn e o desconhecimento da matriz de custos e benefícios, o cálculo do valor esperado, até então avaliado como a melhor métrica para alinhar o projeto com os interesses financeiros do negócio, foi descartado como métrica de verificação de performance neste projeto. Pela sua capacidade de lidar com as incertezas mencionadas, a AUC-ROC foi escolhida como métrica de otimização.

Para o problema que queremos resolver, a forma mais direta de estruturar os dados é com um tabela onde cada linha representa um cliente e cada coluna representa uma das múltiplas informações que podemos usar para caracterizar os clientes. Como os dados fornecidos encontram-se no formato de séries temporais, onde cada linha representa um pedido realizado por algum dos clientes, torna-se necessário reestruturar os dados em um formato mais adequado para a modelagem, extraindo no processo características descritivas (features) das séries temporais.

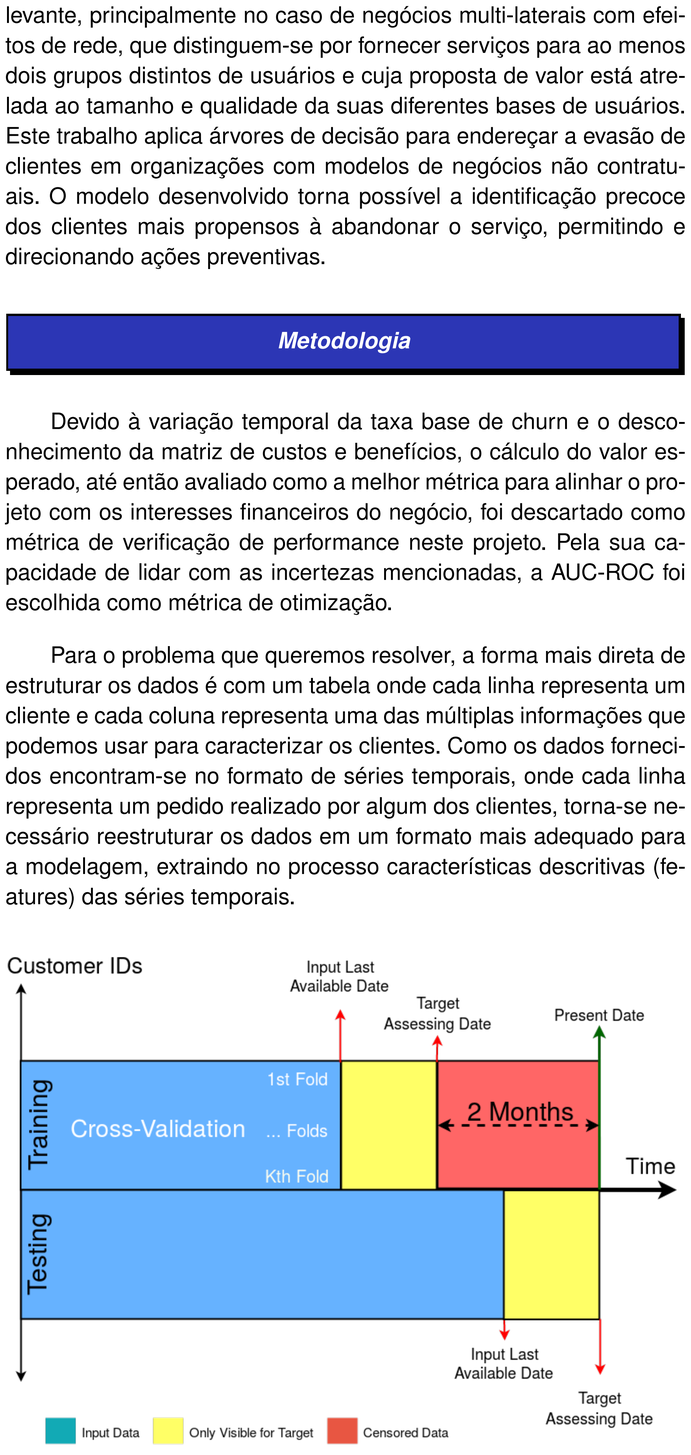

Figura 1: Esquema de Validação

A característica não-contratual do negócio faz com que a definição do target seja um desafio por si só. Para fins de simplicidade, adotou-se a seguinte definição de cliente: Um cliente é considerado ativo se ele tiver feito ao menos um pedido nos últimos 30 dias e inativo caso contrário.

Para criar um modelo verdadeiramente preditivo é necessário prevenir que informações sobre eventos futuros sejam utilizadas no treinamento do mesmo. Por esta razão, parte dos dados foi censurada no cálculo das variáveis de entrada de modo a representar o atraso de 30 dias existente na determinação do target.

As features foram extraídas de forma tanto manual, a partir do próprio conhecimento de negócios e do que foi encontrado na literatura, quanto automatizada, a partir da biblioteca python *tsfresh*, apropriada para a caracterização de séries temporais.

Os dados disponíveis foram separados em dois subsets: um usado para treinar o modelo e outro para testar sua perfomance. Dessa forma, é possível identificar quão bem o modelo lida com dados não vistos. A separação foi feita tanto pelos IDs dos clientes, para que o mesmo cliente não apareça simultaneamente nos dois subsets, quanto temporalmente, privando o modelo de aprender com informações dos 2 últimos meses disponíveis na base de dados. Isto é necessário pois a passagem do tempo pode interferir em características relevantes dos dados e testar a performance do modelo em dados coletados cronologicamente próximos nos daria uma ideia falsamente otimista da capacidade preditiva do modelo.

Resultados e Discussão

Dentre as quase 400 features extraidas das séries temporais analisadas, somente 6 foram identificadas como relevantes. Apesar disso, mais de 91% da redução total da impureza foi obtida por somente duas features, extraídas manualmente: a *recência* e *frequência de pedidos*.

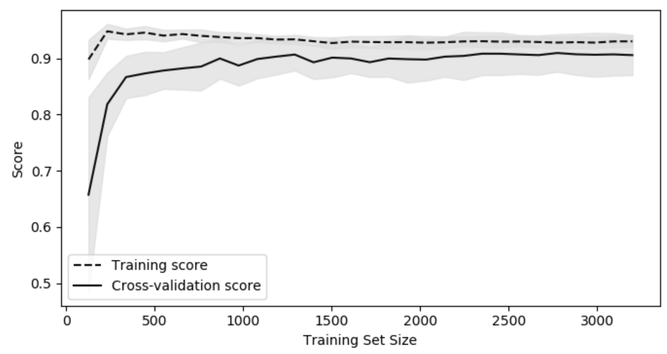

Figura 2: Performance vs Quantidade de amostras

Ao observar a curva de aprendizado (figura 2) podemos dizer que a quantidade de dados fornecida aumenta a capacidade preditiva do modelo de forma crescente até um limite localizado após aproximadamente 1000 amostras.

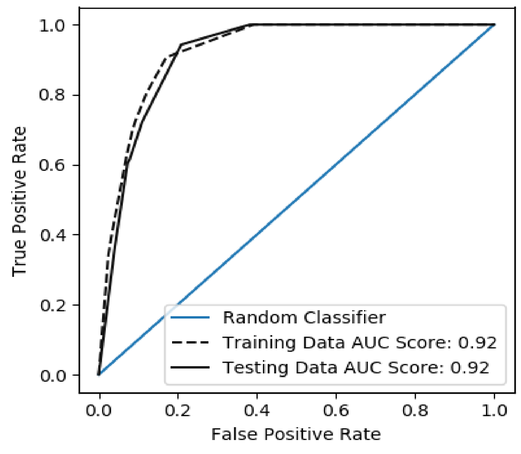

Figura 3: Receiver Operating Characteristics

Além disso, a performance está bem próxima do classificador ideal (AUC = 1.0) e a variação média de performance entre treino e teste tem valor desprezível, indicando que o modelo está com sua complexidade ajustada corretamente e possui níveis aceitáveis de viés e variância.

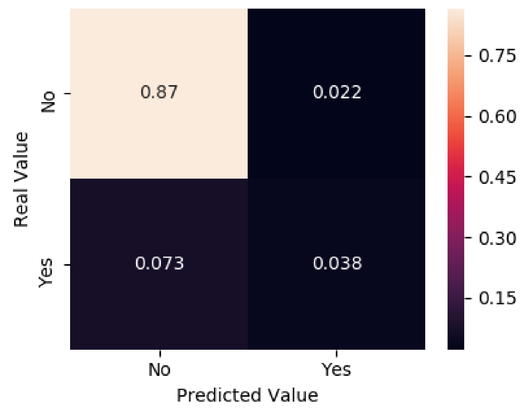

Figure 4: Matriz de Confusão do modelo encontrado

A matriz de confusão (figura 4) nos ajuda a compreender que tipo de erro o modelo de classificação está cometendo, separando em diferentes células cada par combinatório de valor real e previsto. A partir dos números apresentados podemos facilmente calcular a taxa base de churn presente nos dados utilizados (próximo de 11%), a precisão e a sensibilidade do modelo.

$$precision = \frac{TruePositives}{\sum PredictedConditionPositives} \quad (1)$$

$$sensitivity = \frac{TruePositives}{\sum RealConditionPositives} \quad (2)$$

$$f1\_score = \frac{2 * precision * sensitivity}{precision + sensitivity} \quad (3)$$

Percebemos que somente 34% das amostras classificadas como churn foram marcadas corretamente (sensibilidade) e, dentre todas as amostras marcadas como churn, 63% de fato o são (precisão). A média harmônica das duas (f1 score) nos dá um valor próximo de 44%, longe do ideal 100%. Isso porque este modelo, apesar de razoavelmente preciso, falha em identificar a maioria dos casos de churn (baixa sensibilidade).

Conclusão

A oportunidade de trabalhar em um desafio real de negócios tornou possível aplicar ferramentas de desenvolvimento em alta demanda no mercado e praticar de forma simultanea conhecimentos de negócios, data mining e de engenharia de software. Apesar do pouco tempo disponível para a elaboração e implementação do projeto, o resultado final foi bastante satisfatório, pois cumpriu com a proposta. O modelo, apesar de ainda não estar pronto para ser implantado em ambiente de produção de forma automatizada, já pode ser utilizado para gerar insights e direcionar o fluxo decisório da empresa.

Referências

[1] HAGIU, Andrei; WRIGHT, Julian. **Multi-Sided Platforms**. Harvard Business School, 2015

[2] PYLE, Dorian. **Data Preparation for Data Mining**. San Francisco: Morgan Kaufmann Publishers, 1999

[3] FADER, Peter et al. **"Counting your Customers"the Easy Way:** An Alternative to the Pareto/NBD Model, 2005

[4] PROVOST, Foster; FAWCETT, Tom. **Data Science for Business**: What You Need to Know About Data Mining and Data-Analytics Thinking. Sebastopol: O'Reilly Media, 2013.