Modelos de regressão com fração de cura em sobrevivência, aplicado na classificação de clientes com dados segmentados

Douglas de Paula Nestlehner **Orientador: José Carlos Fogo**

Trabalho de Conclusão de Curso - Bacharelado em Estatística

Resumo

Modelos com fração de cura permitem considerar indivíduos que não foram sujeito ao evento de interesse, podendo assim trazer inferências mais precisas em relação aos usuais modelos de sobrevivência. Apresentamos nesse estudo as vantagens e desvantagens obtidas ao se aplicar modelos de sobrevivência com fração de cura para dados segmentados, realizando ajustes considerando diferentes distribuições, no intuito de se obter o melhor modelo, aplicando os estudos em um problema de fidelização de clientes.

1. Introdução

Uma das características que frequentemente observamos em dados coletados no intuito de se realizar algum estudo de sobrevivência, é o alto número de não presença do evento de interesse, o qual denotamos como censura. Em estudos de fidelização de clientes de serviços por assinatura, teremos dados com essa característica, pois em serviços por assinaturas os clientes tendem a consumir o produto por um longo período de tempo.

Diversos fatores vêm fazendo com que clientes de serviços por assinatura (streaming, tv a cabo, internet, etc.) cancelem seus planos e/ou troquem por outros, sendo os principais:

- Os efeitos econômicos da pandemia, que causam a necessidade dos clientes procurarem por serviços mais baratos e alternativos;
- O aumento da concorrência, com cada vez mais novas empresas surgindo no mercado, oferecendo produtos similares e com preços menores;
- A facilidade adquirida pelas tecnologias atuais (internet, telefone, serviços de comunicação etc.) onde conseguimos ter conhecimento dos melhores produtos disponíveis, permitindo a troca de produto quando não estamos satisfeitos;

Portanto, é cada vez mais valorizado a fidelização de clientes, pois é de extrema importância para as empresas de assinaturas ter clientes que continuam adquirindo e utilizando os seus produtos.

2. Objetivos

Neste trabalho iremos utilizar modelos de sobrevivência com fração de cura, na modelagem de dados segmentados, aplicados em um estudo de ocorrência de abandono de serviço por assinatura.

Mais especificamente, o presente trabalho tem como objetivo utilizar modelos de regressão com fração de cura no intuito de predizer a probabilidade de um cliente de serviços por assinatura, permanecer fiel por um longo período de tempo, considerando suas características.

3. Materiais e Métodos

Ao longo do trabalho, estudamos os principais conceitos de analise de sobrevivência, sendo os principais:

3.1 Censura

- **Censura do Tipo I:** Ocorre quando o estudo tem um limite final de tempo de execução (denotado por L), pré-definido, e quando o tempo L é atingido, todos os indivíduos que não apresentaram o evento de interesse são considerados como censura do tipo I.
- **Censura do Tipo II:** Ocorre quando o estudo tem um limite pré-definido de eventos de interesse, ou seja, em uma amostra com n indivíduos é observada até a ocorrência de r eventos de interesse ($r < n$), assim os $n - r$ indivíduos restantes são considerados como censura.
- **Censura Aleatória:** Ocorre quando o indivíduo por algum motivo aleatório deixa o estudo. Observada com frequência na área médica, quando um paciente deixa o estudo por razões distintas do evento de interesse, como exemplo: morte do paciente por outras razões não relacionadas ao estudo; paciente deixa de comparecer ao estudo; etc.

3.2 Modelo com Fração de Cura

Para que esse tipo de modelo seja uma opção viável é necessário que o tempo de seguimento dos indivíduos seja suficientemente longo, e a curva de sobrevivência apresente uma estabilização a partir de um determinado tempo, indicando a presença de uma fração razoável de indivíduos que não irão experimentar o evento de interesse, mesmo se forem acompanhados por um longo período de tempo.

Da modelagem com fração de cura, a função de sobrevivência $S(t)$ é conhecida como própria, porém, para os indivíduos que não apresentam o evento de interesse são considerados como tendo tempo de vida infinito.

Desta forma, a probabilidade do tempo de vida ser maior do que um valor t é definida por $S_{pop}(t)$ (sobrevivência populacional ou imprópria), sendo dada por:

$$S_{pop}(t) = (1 - p) + pS(t),$$

em que p é a probabilidade do individuo estar em risco no instante t.

Sob esse enfoque, a fração de indivíduos que não apresentam o evento de interesse, também chamada de fração de cura, é dada por $(1 - p)$.

Sendo a densidade e função de risco impróprias dadas por:

$$f_{pop}(t) = pf(t),$$
$$h_{pop}(t) = \frac{pf(t)}{(1-p) + pS(t)}.$$

Dessa forma, a função de verossimilhança para o modelo com fração de cura, é dada por:

$$L(\theta|D) = \prod_{i=1}^{n} \left[f_{pop}(t)\right]^{\delta_i} \left[S_{pop}(t)\right]^{1-\delta_i}$$
$$= \prod_{i=1}^{n} \left[pf(t)\right]^{\delta_i} \left[(1-p) + pS(t)\right]^{1-\delta_i}$$

Considerando g grupos, cada um com uma fração de cura p_i, com $i = 1, 2, ..., g$, a função de verossimilhança para o i-ésimo grupo, é dada por:

$$L_i(\theta|D) = \prod_{j=1}^{n_i} \left[p_i f(t)\right]^{\delta_{ij}} \left[(1-p_i) + p_i S(t)\right]^{1-\delta_{ij}}$$

3.3 Modelo Exponencial

Considerando a distribuição exponencial para os tempos de vida, a função de densidade $f(t)$ é dada por:

$$f(t_{ij}) = \lambda \mathrm{e}^{-\lambda t_{ij}}, \quad i = 1, 2, ..., g \quad \text{e} \quad j = 1, 2, ..., n_i.$$

Teremos então, a sobrevivência imprópria do j-ésimo individuo do i-ésimo grupo, dada por:

$$S_{pop}(t_{ij}) = (1 - p_i) + p_i \mathrm{e}^{-\lambda t_{ij}},$$

e a densidade imprópria, por:

$$f_{pop}(t_{ij}) = \lambda p_i \mathrm{e}^{-\lambda t_{ij}},$$

Assim, a função de verossimilhança do i-ésimo grupo é dada por:

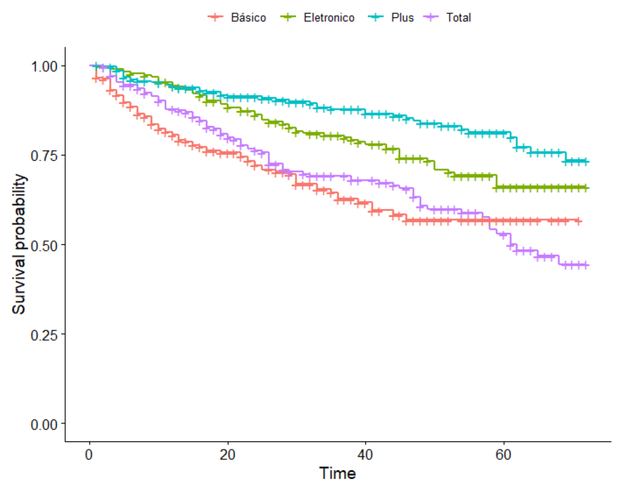
$$L_i(\theta|D) = \prod_{j=1}^{n_i} \left[\lambda p_i \mathrm{e}^{\lambda t_{ij}}\right]^{\delta_{ij}} \left[(1 - p_i) + p_i \mathrm{e}^{-\lambda t_{ij}}\right]^{1-\delta_{ij}}$$

4. Resultados

Para a aplicação, utilizamos dados empresa Telecom, os quais foram captados no intuito de obter informações sobre $churn$ de seus clientes. O experimento teve uma duração de 72 meses, nos quais foram coletadas informações de 1000 clientes por meio de 20 covariáveis.

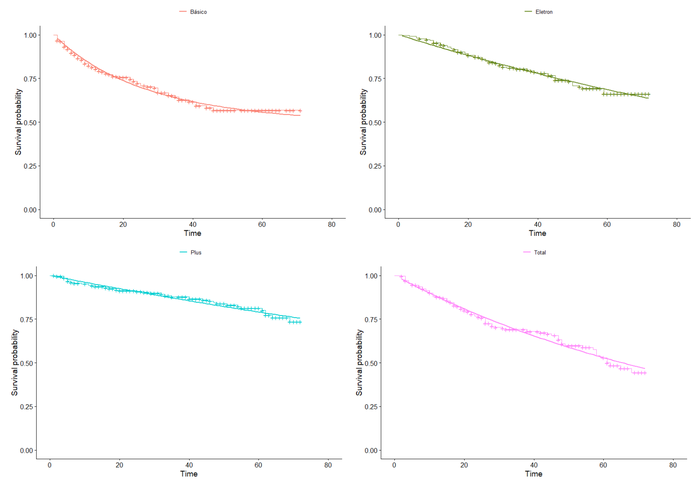
	Básico	Eletrônico	Plus	Total
Ocorrência $churn$	31.20%	27.19%	15.66%	37.29%
Censura	68.80%	72.81%	84.34%	62.71%

No intuito de observar o comportamento da função de sobrevivência para as diferentes segmentações, utilizamos o estimador de Kaplan-Meier, que nos permite identificar ou não a fração de cura.



Conseguimos notar apenas a segmentação “Básico” tem uma “cauda” longa, o que significaria uma concentração de censura maior ao fim do experimento (censura do tipo I), o que caracteriza a fração de cura.

Apesar de observarmos pelo Kaplan-Meier que a aplicação de modelos com fração de cura pode não ser adequado para algumas segmentações, ajustamos um modelo com fração de cura considerando a distribuição exponencial, no intuito de observar o comportamento do ajuste.



Notamos que aparentemente as estimativas são boas, ou seja, as estimativas obtidas pelos modelos se aproximam de seus respectivos Kaplan-Meier. Entretanto, é importante observarmos os parâmetros estimados na modelagem, pois assim conseguimos identificar o real desempenho dos modelos, e observar a fração de cura estimada para cada segmentação.

Parâmetro	Segmentação	Estimativa	IC (95%)
p	Básico	0.4819	(0.3280 - 0.6337)
p	Eletrônico	0.0039	(5.31e-23 - 1.00e+00)
p	Plus	0.0068	(9.23e-24 - 1.00e+00)
p	Total	0.0015	(7.61e-22 - 1.00e+00)
λ	Básico	0.02693	(0.0140 - 0.05157)
λ	Eletrônico	0.0063	(4.52e-03 - 8.79e-03)
λ	Plus	0.0039	(2.46e-03 - 6.29e-03)
λ	Total	0.0106	(8.49e-03 1.33e-02)

As estimativas das probabilidades $(1-p)$ dos clientes continuarem fiéis ao serviço por assinatura, para as diferentes segmentações, são estimativas ruins em comparação as porcentagens apresentadas na Tabela 1. A explicação para esses resultados se deve principalmente ao que foi observado no Kaplan-Meier, o conjunto de dados é caracterizado por apresentar dados com censuras do tipo aleatórias,

5. Considerações Finais

De modo geral, com o problema encontrado na aplicação, conseguimos notar diversos pontos de destaques:

- Devemos sempre investigar os tipos de censura presente nos dados em questão, permitindo assim, verificar se aplicação de modelos com fração de cura é viável.
- Apesar de não conseguirmos interpretar os resultados da fração de indivíduos curados, o ajuste do modelo com fração de cura consegue estimar a curva de sobrevivência dos indivíduos não curados (ocorrência de $churn$), permitindo a interpretação dos resultados, assim como é feito nos usuais modelos de sobrevivência;
- A segmentação dos dados em características importantes, podem trazer resultados relevantes para cada grupo, permitindo assim, traçar estratégias mais especificas.

Para a próxima aplicação, pretendemos ajustar modelos de sobrevivência com fração de cura considerando as distribuições exponecial, Weibull e Gompertz, e comparar o desempenho das mesmas. Se possível, também pretendemos estudar outros tipos de modelagem de fração de cura, permitindo a comparação com a modelagem da fração de cura apresentada nesse trabalho.

Referências

[1] Maller, R. A. e Zhou, X. *Survival analysis with long-term survivors*, New York: Wiley, 1996.

[2] Calsavara, V. F. *Modelos de sobrevivência com fração de cura usando um termo de fragilidade e tempo de vida Weibull modificada generalizada*, Dissertação de Mestrado, Universidade Federal de São Carlos, 2011.

[3] Fogo, J. C. *Modelo de regressão para um processo de renovação Weibull com termo de fragilidade*, Tese (Doutorado em agronomia), Escola superior de Agronomia Luiz de Queiroz, Universidade de São Paulo, 2007.

[4] Collet, D. *Modelling survival data in medical researchem*, ICRC press, 2015.