

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

Everson Luiz Henequi

# **Modelo de Projeção da Taxa de Pagamento em Cobrança**

**Curitiba  
2019**

Everson Luiz Henequi

## **Modelo de Projeção da Taxa de Pagamento em Cobrança**

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. José Luiz Padilha da Silva

Curitiba  
2019

# Modelo de Projeção da Taxa de Pagamento em Cobrança

Aplicação da biblioteca GAMLSS para criação de modelos sobre variáveis resposta contínuas e inflacionadas em 0 e 1

Everson Luiz Henequi<sup>1</sup>

<sup>1</sup>Aluno do programa de Especialização em Data Science & Big Data \*

## Resumo

Este artigo oferece uma abordagem alternativa à regressão logística tradicional para gerenciamento de risco de crédito, especificamente para cobranças de clientes inadimplentes. É comum tratar esse problema com uma abordagem dicotômica entre os clientes que pagam suas dívidas e os que não pagam. No entanto, o dinheiro coletado é muitas vezes ignorado pelo processo de modelagem, considerando-se 1% e 100% do pagamento de dívidas igualmente. Portanto, há uma oportunidade para estudar o percentual esperado de pagamento da dívida. Criar um modelo que forneça tal informação ajudará muito na melhoria, tanto da estratégia de abordagem aos clientes, quanto dos indicadores de desempenho. Pretende-se suportar as decisões gerenciais com um modelo preditivo, utilizando o *framework* GAMLSS, que forneça o percentual esperado de recebimento da dívida dado o perfil comportamental e demográfico do cliente.

**Palavras-chave:** Analytics, Cobrança, Banco, GAMLSS, Recuperação de Crédito, Modelagem

## Abstract

This paper offers an alternative approach to the traditional logistic regression for credit risk management, specifically for delinquent customer collections. It is common to treat this problem with a dichotomous approach between customers who pay their debts and those who do not pay. However, the cash collected is often ignored by modeling process, considering 1% and 100% of the debts equally. Therefore, there is an opportunity to study the expected percentage of debt repayment. Creating a model that provides such information will greatly help in improving both the recovery strategy and the performance indicators. With this, it is intended to support managerial decisions with a predictive model, using *GAMLSS framework*, that provides the expected percentage of debt receipt given the behavioral and demographic profile of the client.

**Keywords:** Analytics, Collections, Bank, GAMLSS, Credit Recovery, Modeling

## 1. Introdução

Para uma instituição financeira, conceder crédito implica obrigatoriamente em controlar o nível da incerteza do retorno financeiro, bem como dos juros remuneratórios pelo risco assumido. Por se tratar de uma relação com a presença de risco, alguns dos consumidores não são capazes de honrar suas obrigações e são categorizados como inadimplentes. O desafio está portanto, em identificar clientes potenciais inadimplentes, ditos como “maus clientes”, diferenciando-os daqueles que pagarão suas dívidas, considerados “bons”. De posse desta informação é possível otimizar esforços e reduzir custos com o processo de retomada do montante em risco. As abordagens de modelagem comumente utilizadas para a predição dicotômica

descrita acima, contudo na prática diária de cobrança de inadimplentes, a negociação com os “bons clientes” resulta em um acordo entre as partes, onde há a aplicação de redutores da dívida (descontos) para baixa do registro de inadimplência. Neste caso os bons clientes podem ter suas pendências resolvidas pagando um percentual da dívida que varia entre 1% e 100% da dívida.

Empiricamente, do ponto de vista de negócios, é uma vantagem competitiva o cálculo apurado da recuperação esperada para cada dívida. A partir desta informação, a oferta de negociação pode ser mais precisa e pode-se ter mais agilidade nas decisões de venda de carteira com baixíssima expectativa de recuperação. Por isto, o presente estudo pretende estabelecer um modelo capaz de projetar o percentual esperado de pagamento da dívida de clientes inadimplentes de um

\*henequi@gmail.com

grande banco com atuação nacional, a partir de uma série de características comportamentais históricas e demográficas. Especificamente trabalhou-se com clientes cujas dívidas estão em atraso de 90 a 120 dias, por tratar-se de uma população onde é mais viável ações de cobrança, tais como: campanhas de recuperação, meritocracia e calibração da oferta de negociação.

Dado o contexto estabelecido, aliado a característica de concentração de clientes que pagam 100% da dívida e principalmente uma grande concentração em clientes que não pagam sua dívida (0%), realizou-se uma modelagem estatística utilizando o arcabouço de técnicas do *framework* GAMLSS (modelos aditivos generalizados para locação, escala e forma).

## 2. Conceituação e definições

### 2.1. População Alvo

Os dados utilizados no desenvolvimento do estudo têm como fonte principal bases de inadimplentes com atraso entre 90 e 120 dias em produtos de conta corrente, financiamentos e cartões de crédito.

A seleção dos meses para o desenvolvimento e validação do modelo seguiu uma abordagem comumente utilizada em modelagem estatística de dados, onde capturam-se vários meses do ano para mitigar o risco de influência sazonal do modelo. As safras escolhidas foram: Desenvolvimento: Fev, Mai, Ago e Nov/2017 Validação: Dez/2017, Jan e Fev/2018

**Tabela 1:** Quantidade das Populações Alvo

|         | Mês      | Quantidade |
|---------|----------|------------|
| Desenv. | Fev/17   | 12.965     |
|         | Mai/17   | 14.711     |
|         | Ago/17   | 11.283     |
|         | Nov/17   | 9.549      |
|         | Subtotal | 48.508     |
| Valid.  | Dez/17   | 39.835     |
|         | Jan/18   | 40.168     |
|         | Fev/18   | 39.019     |

Para evitar que algum cliente se repita na base de desenvolvimento, foi atribuído um valor aleatório (entre 00 e 99) para cada cliente e coletou-se 25% de cada safra, por exemplo para o Mar/17 selecionou-se somente os dígitos de 00 a 24, para Ago/17 os dígitos 25 a 49 e assim por diante. Este processo garantiu a unicidade

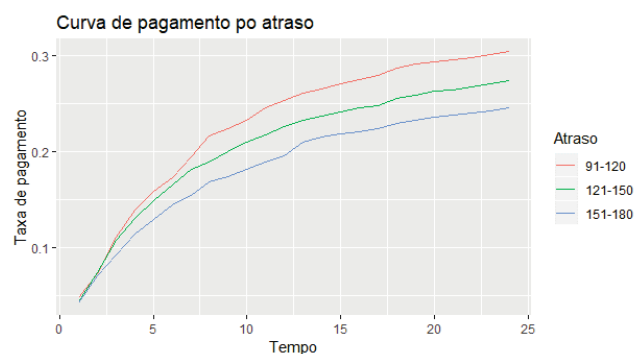
do cliente e tornou o processo computacional mais ágil, por este motivo as quantidades da amostra de desenvolvimento são inferiores aos meses de validação.

### 2.2. Definição da variável resposta

O problema que pretende-se estudar aqui refere-se ao pagamento de dívidas a partir de um determinado ponto de observação. Com isto, algumas características devem ser observadas para definir assertivamente a variável resposta, são elas:

**Pagamento à vista** - quando o cliente quita sua dívida em parcela única, neste caso normalmente a negociação pode incorrer em um desconto maior.

**Renegociação de dívida** - nesta modalidade, na impossibilidade saldar a dívida na modalidade anterior, o cliente aceita um acordo onde ocorre a substituição do empréstimo anterior por outro com melhores condições de pagamento, geralmente alongando prazo e reduzindo o valor da parcela.



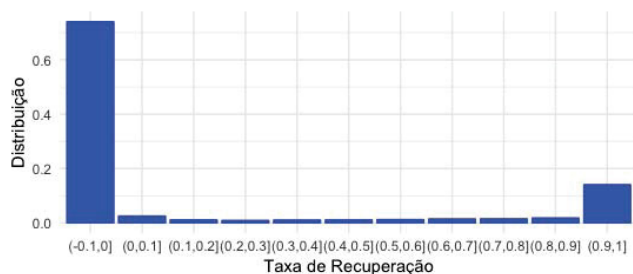
**Figura 1:** Percentual de pagamento de dívidas acompanhados em um horizonte de 24 meses.

A Figura 1, demonstra o comportamento de populações similares ao público de interesse e foi observado numa janela de performance de 24 meses. Percebe-se uma maior incidência de pagamentos no de curto prazo.

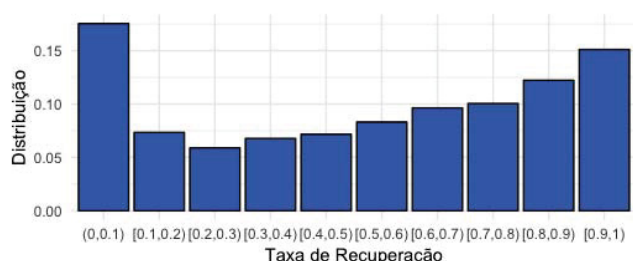
Os processos rotineiros de cobrança chegam a levar em conta 5 a 10 anos a partir do ponto de observação, isto representa um período muito longo onde as covariáveis não se mostram discriminantes, por isso fez-se uma pesquisa específica quanto a estabilidade e força de predição das variáveis disponíveis em relação a taxa de pagamento acumulada. Com isto chegou-se a uma definição de que a janela de performance de 6 meses, após o ponto de observação, seria adequada para este estudo.

Outro requisito importante a ser levado em consideração é o tempo até o cliente aceitar uma das duas modalidades acima, em outras palavras o cliente pode aderir ao pagamento em qualquer ponto entre o início e o fim do período de performance.

A partir das definições acima, foi possível avaliar como a população se distribui quanto ao objeto de interesse do estudo. A Figura 2 abaixo apresenta claramente uma concentração nos clientes que não pagam suas dívidas (0% de taxa de pagamento), bem como uma concentração significativa nos clientes que saldam 100% de suas dívidas. Adicionalmente na Figura 3 demonstra-se a distribuição da população excluindo-se os percentuais 0% e 100%.



**Figura 2:** Distribuição populacional em relação do percentual de pagamento



**Figura 3:** Distribuição populacional em relação do percentual de pagamento (Exceto 0 e 1)

### 2.3. Variáveis Covariáveis

De posse das definições acima, fomos em busca de informações que possam nos servir de componentes preditoras para o modelo. Com isto, coletou-se dados em todas as fontes disponíveis, entre elas:

- Restritivos de Crédito
- Dados Cadastrais
- Histórico Pagamentos / Atrasos
- Histórico de Relacionamento com o Banco
- entre outras

Após intenso processo de agregação atingiu-se um número de 830 variáveis candidatas a compor o modelo final.

## 3. Metodologia

### 3.1. Tratamento dos Dados

Após as primeiras exclusões de variáveis com preenchimento inferior a 1%, foi realizada uma análise minuciosa, variável a variável, para identificar e tratar possíveis problemas nas mesmas. As variáveis numéricas com comportamento assimétrico foram transformadas utilizando a função logarítmica na base  $e$  (2,71828...) ou na base 10, dependendo do nível da assimetria. Variáveis com comportamento simétrico porém com escalas muito grandes foram divididas pelos seus respectivos desvios padrões. Existem ainda variáveis cujos valores ausentes não são devido à perda de informação e sim à não exposição à uma determinada característica, neste caso os valores ausentes foram transformados em uma categoria da variável e caso o restante dos valores da mesma fossem contínuos seriam também transformados em categorias utilizando alguns pontos de corte. Um exemplo deste último caso é quando tem-se uma variável indicadora do número de restritivos que foram baixados do mercado, ou seja, quantos dos restritivos que um cliente possuía não existem mais, neste caso a variável pode assumir valores entre 0 e o número de restritivos que o cliente possuía, no entanto existe o valor ausente que significa que o cliente nunca teve um restritivo para que o mesmo pudesse ser baixado.

Por fim, para algumas variáveis categóricas observou-se categorias com baixa frequência, as quais foram agrupadas com a categoria adjacente de forma iterativa até que possuíssem uma frequência significativa.

### 3.2. Imputação de Dados Faltantes

A partir deste ponto o ajuste de modelos contendo todas as variáveis torna-se inviável caso não haja um tratamento nos dados ausentes uma vez que na maioria das técnicas de modelagem não são ajustados caso uma linha tenha pelo menos um dado ausente. A exclusão dessas linhas não é viável e pode viesar o modelo, pois observações que apresentam valores ausentes podem ter uma relação distinta com a variável resposta. Neste caso optou-se por preencher os valores ausentes através do método de imputação implementado pelo pacote *mice*.

A função  $m_i$  é imputa nas covariáveis numérica através do método de correspondência média preditiva que se baseia em modelos lineares da covariável que será imputada utilizando como possíveis covariáveis todas as demais variáveis do conjunto de dados. Para as covariáveis categóricas com 2 fatores a função utiliza o modelos de regressão logística. Para as covariáveis categóricas com mais de 2 fatores não ordenados a função utiliza modelos politômicos. E para as covariáveis categóricas com mais de 2 fatores ordenados a função utiliza modelo de chances proporcionais.

### 3.3. Modelagem do Problema

O ferramental estatístico escolhido para alcançar o objetivo de modelar os dados e em especial o percentual de recuperação dos clientes inadimplentes foi o *framework* GAMLSS (*Generalized Additive Models for Location, Scale and Shape*), ou em uma tradução literal, os modelos aditivos generalizados para localização, escala e forma. A metodologia abordada em GAMLSS permite que os preditores lineares estejam associados não somente a um parâmetro da distribuição, mas a quantos outros a distribuição comportar.

Tratam-se de modelos de regressão semi-paramétricos, sendo paramétricos, na medida em que requerem uma suposição de distribuição paramétrica para a variável de resposta, e “semi” no sentido de que a modelagem dos parâmetros da distribuição, como funções de variáveis explicativas, podem envolver o uso de funções de suavização não paramétricas. GAMLSS foram introduzidos por Rigby e Stasinopoulos (2001, 2005) e Akantziliotou, Rigby e Stasinopoulos (2002) como forma de superar algumas das limitações associadas aos populares modelos lineares generalizados, GLM e modelos aditivos generalizados, GAM (ver Nelder e Wedderburn, 1972; Hastie e Tibshirani, 1990, respectivamente).

Como observado na seção a respeito do conjunto de dados, a variável resposta possui um comportamento atípico, com domínio no intervalo  $[0, 1]$  e com inflação em 0 e 1. Um modelo GAMLSS cuja variável resposta segue uma determinada distribuição denotada por  $D$ , com parâmetros  $\mu, \sigma, \nu$  e  $\tau$ , isto é,  $Y \sim D(\mu, \sigma, \nu, \tau)$ , pode ser expresso da seguinte forma:

$$\begin{aligned} g_1(\mu) &= X_1\beta_1 + s_{11}(x_{11}) + \dots + s_{1J_1}(x_{1J_1}) \\ g_2(\sigma) &= X_2\beta_2 + s_{21}(x_{21}) + \dots + s_{2J_2}(x_{2J_2}) \\ g_3(\nu) &= X_3\beta_3 + s_{31}(x_{31}) + \dots + s_{3J_3}(x_{3J_3}) \\ g_4(\tau) &= X_4\beta_4 + s_{41}(x_{41}) + \dots + s_{4J_4}(x_{4J_4}), \end{aligned} \quad (1)$$

em que  $g_i(\cdot)$  representa a função de ligação,  $X_i$  o vetor de covariáveis associadas ao parâmetro contido função de ligação  $g_i(\cdot)$ ,  $\beta_i$  os coeficientes associados aos parâmetros  $X_i$  e  $S_{ij}(x_{ij})$  a função suavizada da covariável  $x_{ij}$ , sendo  $i = 1, \dots, 4$  e  $j = 1, \dots, J_i$ .

Em GAMLSS, uma distribuição que pode ser utilizada para modelar a variável resposta deste estudo é a distribuição Beta Inflacionada em Zero e Um, trata-se de uma mistura de três componentes: um valor discreto 0, ao qual é associada uma probabilidade  $p_0$ , um valor discreto 1, ao qual é associado uma probabilidade  $p_1$ , e uma distribuição beta  $BE(\mu, \sigma)$  no intervalo  $(0, 1)$  com probabilidade  $p_2 = (1 - p_0 - p_1)$ . A Beta Inflacionada em Zero e Um é definida através de quatro parâmetros,  $\mu$  e  $\sigma$  que definem a distribuição beta  $BE(\mu, \sigma)$  associada à parte contínua da distribuição beta inflacionada, em que  $0 < \mu < 1$  e  $0 < \sigma < 1$ ,  $\nu$  que está associada a inflação de zeros e  $\tau$  que está associada a inflação de uns. As duas componentes associadas às probabilidades de 0 e 1 citadas anteriormente estão relacionadas aos parâmetros da distribuição da seguinte forma:

$$\begin{aligned} p_0 &= \frac{\nu}{1 + \nu + \tau} \\ p_1 &= \frac{\tau}{1 + \nu + \tau} \end{aligned} \quad (2)$$

Deste modo, definiu-se a função densidade de probabilidade da distribuição Beta Inflacionada em Zero e Um da seguinte forma:

$$f_Y(y | \underline{\theta}) = \begin{cases} p_0 & \text{se } y = 0 \\ (1 - p_0 - p_1)f_W(y | \mu, \sigma) & \text{se } 0 < y < 1, \\ p_1 & \text{se } y = 1 \end{cases} \quad (3)$$

em que  $W$  segue uma distribuição beta de parâmetros  $\mu$  e  $\sigma$ .

Para cada parâmetro é ainda associada uma função de ligação, utilizando a logística para  $\mu$  e  $\sigma$  e a logarítmica com base  $e$  para  $\nu$  e  $\tau$ .

Neste estudo tem-se especial interesse nos parâmetros  $\mu, \nu$  e  $\tau$ . Para fins de aplicação prática não tratou-se de  $\sigma$  que fornece informação da variância da estimativa de  $\mu$ .

### 3.4. Seleção das Covariáveis

A metodologia GAMLSS permite que uma covariável esteja relacionada a um ou mais parâmetros da distribuição, fato este que nos permitiu avaliar o desempenho

das covariáveis através do ajuste de modelos definidos de diferentes formas. Cada covariável foi inserida em um modelo *GAMLSS* de forma isolada, ou seja, não ajustada pelas demais variáveis, deste modo, para cada variável foi possível criar 8 modelos incluindo o modelo nulo (modelo em que a covariável não está associada a nenhum dos parâmetros da distribuição). As especificações desses 8 modelos se diferem com relação aos parâmetros relacionados à covariável. A Tabela 2 especifica os 8 modelos, em que “✓” indica a presença da variável no parâmetro.

**Tabela 2:** Especificação dos modelos

| Modelos | Parâmetros |          |       |        |
|---------|------------|----------|-------|--------|
|         | $\mu$      | $\sigma$ | $\nu$ | $\tau$ |
| $y_0$   |            |          |       |        |
| $y_1$   | ✓          |          |       |        |
| $y_2$   |            |          | ✓     |        |
| $y_3$   |            |          |       | ✓      |
| $y_4$   | ✓          |          | ✓     |        |
| $y_5$   | ✓          |          |       | ✓      |
| $y_6$   |            |          | ✓     | ✓      |
| $y_7$   | ✓          |          | ✓     | ✓      |

Para esta etapa o parâmetro  $\sigma$  não foi utilizado. Cada variável irá se submeter ao ajuste dos 8 modelos, isso inclui o modelo nulo pois o mesmo será ajustado de acordo com a disponibilidade de dados da variável que está sendo considerada, ou seja, excluindo-se os dados ausentes.

Dentre as possíveis medidas para avaliar o ajuste do modelo, testou-se o AIC (Critério de Informação de Akaike) e o BIC (Critério de Informação Bayesiano) que penaliza a verossimilhança de acordo com o número de observações. Essas medidas podem ser expressa da seguinte forma:

$$\begin{aligned} AIC &= 2\ln(L) + 2[(p + 1) + 1] \\ BIC &= 2\ln(L) + \ln(n)[(p + 1) + 1], \end{aligned} \quad (4)$$

em que  $L$  é a máxima verossimilhança do modelo e  $p$  é o número de parâmetros.

A variável que apresentou o menor AIC ou BIC no modelo 1 (modelo nulo) foi retirada da análise pois, em teoria, a presença da mesma no modelo apenas piorou o ajuste, o que nos faz aceitar a hipótese de que a mesma não contribuiria para o modelo ajustado pelas demais covariáveis.

A Tabela 1 sugere que algumas covariáveis sejam ajustadas somente para o parâmetro  $\mu$ , outras para  $\nu$ , outras para  $\tau$  e outras para uma combinação desses 3. Deste modo tem-se dois possíveis ajustes iniciais: 1.

ajustar um modelo em que todas as covariáveis que não tiveram o mínimo AIC ou BIC no modelo 1 estejam presentes em todos os parâmetros da distribuição (exceto o parâmetro  $\sigma$ ); 2. ajustar um modelo em que cada covariável estará associada somente aos parâmetros presentes no modelo em que a mesma apresentou o mínimo AIC ou BIC.

Para a seleção de variáveis o pacote *GAMLSS* implementado no R disponibiliza as funções *stepGAICAll.A* e *stepGAICAll.B*, na primeira é possível obtermos como resultado diferentes covariáveis associadas a diferentes parâmetros, na segunda todas as variáveis selecionadas devem estar associadas a todos os parâmetros da distribuição.

Dado as possíveis definições de modelos e os diferentes métodos de seleção de variáveis utilizadas nos mesmos, separou-se os ajustes dos modelos finais em 3 possíveis abordagens: 1. Utiliza o segundo ajuste inicial e o método implementado na função *stepGAICAll.A* com penalização através do BIC, utilizando o argumento  $k = \log(n)$ ; 2. Utiliza o primeiro ajuste inicial e o método implementado na função *stepGAICAll.B* com penalização através do BIC; 3. Utiliza o primeiro ajuste inicial e o método implementado na função *stepGAICAll.A*.

Como o objetivo deste estudo é estritamente preditivo, a avaliação dos modelos foi realizada utilizando as bases de validação de dezembro de 2017 e janeiro e fevereiro de 2018. Com as previsões obtidas através da função densidade de probabilidade (Equação 3) apresentada nesta seção, calculou-se o Erro Quadrático Médio de Predição, dado por:

$$EQMP(\hat{\theta}) = \frac{\sum_{i=1}^n (\hat{\theta} - \theta)^2}{n}, \quad (5)$$

em que  $\hat{\theta}$  é o valor predito na escala da variável resposta,  $\theta$  é o verdadeiro valor da variável resposta e  $n$  é o número de observações. Por fim, elegeu-se o modelo com menor EQMP.

#### 4. Resultados

Inicialmente, tem-se uma quantidade excessiva de potenciais covariáveis e realizou-se uma série de análises descritivas para verificar quais delas apresentavam pouca variância, em seguida buscou-se aquelas variáveis que apresentavam 100% de correlação com outras variáveis, desta forma verificou-se a consistência e unicidade das mesmas, reduzindo significativamente o

número inicial de 830 para 81 covariáveis. Como critério estatístico de seleção, performou-se um processo recursivo de geração de modelos utilizando cada uma das covariáveis isoladamente, para estimar todas as possíveis combinações dos parâmetros  $\mu$ ,  $\nu$  e  $\tau$ . De posse desses modelos calculou-se o BIC (*Bayesian Information Criterion*) dos mesmos e comparou-se com o BIC do modelo nulo (que considera somente o intercepto). Portanto conforme demonstrado na Tabela 3) se o menor BIC observado dentre estes 8, for o do modelo nulo, a covariável poderia ser excluída pois não demonstra relação com o fenômeno estudado e caso contrário esta covariável poderia participar do cálculo de estimativa do(s) parâmetro(s) correspondente(s).

**Tabela 3:** Tabela dos menores BICs por covariável

| Modelos                            | Covariáveis |    |    |    |     |     |
|------------------------------------|-------------|----|----|----|-----|-----|
|                                    | V1          | V2 | V3 | V4 | ... | V84 |
| $y_0 \sim 1$                       |             |    |    |    |     |     |
| $y_1 \sim BeInf(\mu, \cdot)$       |             |    |    |    |     |     |
| $y_2 \sim BeInf(\cdot, \nu)$       |             |    |    |    |     |     |
| $y_3 \sim BeInf(\cdot, \tau)$      |             |    |    |    |     |     |
| $y_4 \sim BeInf(\mu, \nu, \cdot)$  |             |    |    |    |     |     |
| $y_5 \sim BeInf(\mu, \cdot, \tau)$ |             |    |    |    |     |     |
| $y_6 \sim BeInf(\cdot, \nu, \tau)$ |             |    |    |    |     |     |
| $y_7 \sim BeInf(\mu, \nu, \tau)$   |             |    |    |    |     |     |

Abaixo a Tabela 4 traz o detalhamento no número de covariáveis associadas a cada parâmetros.

**Tabela 4:** Número de covariáveis com AIC e BIC mínimos em cada especificação de modelo

|     | Nulo | Modelos |       |        |              |               |               |                    |
|-----|------|---------|-------|--------|--------------|---------------|---------------|--------------------|
|     |      | $\mu$   | $\nu$ | $\tau$ | $(\mu, \nu)$ | $(\mu, \tau)$ | $(\nu, \tau)$ | $(\mu, \nu, \tau)$ |
| AIC | 0    | 1       | 9     | 0      | 15           | 5             | 5             | 46                 |
| BIC | 6    | 2       | 35    | 2      | 8            | 6             | 4             | 18                 |

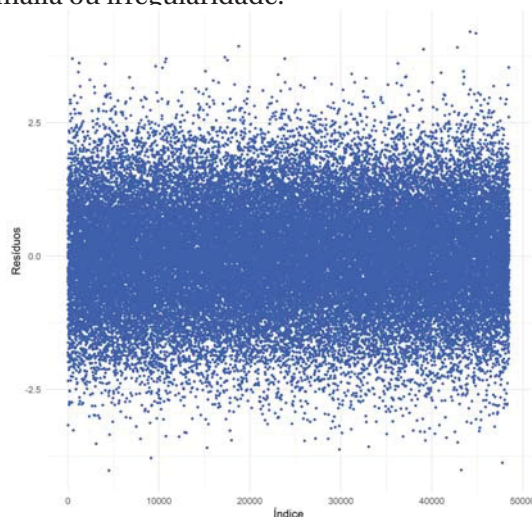
Obteve-se então a 75 covariáveis candidatas a integrar o modelo final. O próximo passo foi submeter as covariáveis selecionadas para modelagem GAMLSS, abaixo na Tabela 5 adotou-se dois tipos de ajustes, o primeiro (Ajuste A), utilizando a seleção de covariáveis da Tabelas 3 e 4 nos parâmetros que melhor BIC apresentou, inclusive podendo ser associada a mais de um parâmetro. Já o Ajuste B incluiu-se inicialmente as 75 covariáveis para cada um dos parâmetros e a partir disso evoluiu-se para o Modelo 2 associando a covariável a todos os parâmetros ou a nenhum, enquanto no Modelo 3 deixou-se livre as covariáveis entrarem ou não na estimativas dos parâmetros.

**Tabela 5:** Número de covariáveis associadas a cada parâmetro e BIC dos Ajustes A e B e Modelos 1, 2 e 3

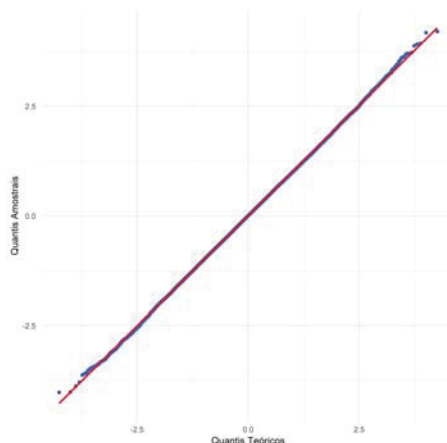
| Modelo   | Parâmetros |       |        | BIC       |
|----------|------------|-------|--------|-----------|
|          | $\mu$      | $\nu$ | $\tau$ |           |
| Ajuste A | 34         | 65    | 30     | 64.385,84 |
| Modelo 1 | 7          | 23    | 10     | 63.900,50 |
| Ajuste B | 75         | 75    | 75     | 65.199,29 |
| Modelo 2 | 22         | 22    | 22     | 64.101,16 |
| Modelo 3 | 8          | 24    | 7      | 63.700,58 |

Ainda sobre a Tabela 5, observou-se uma redução no BIC nos modelos otimizados em relação aos modelos referência, o Modelo 3 apresentou maior redução, sendo aqui a escolha até o momento e segue para as demais verificações e comparações a seguir.

As figuras 4 e 5 apresenta-se os resultados da análise de resíduos quantílicos não apresentando nenhuma anomalia ou irregularidade.



**Figura 4:** Gráfico de Resíduos



**Figura 5:** Gráfico de QQPlot

Portanto, a Tabela 6 apresenta os coeficientes no modelo obtido.

**Tabela 6:** Tabela com os coeficientes do modelo GAMLSS

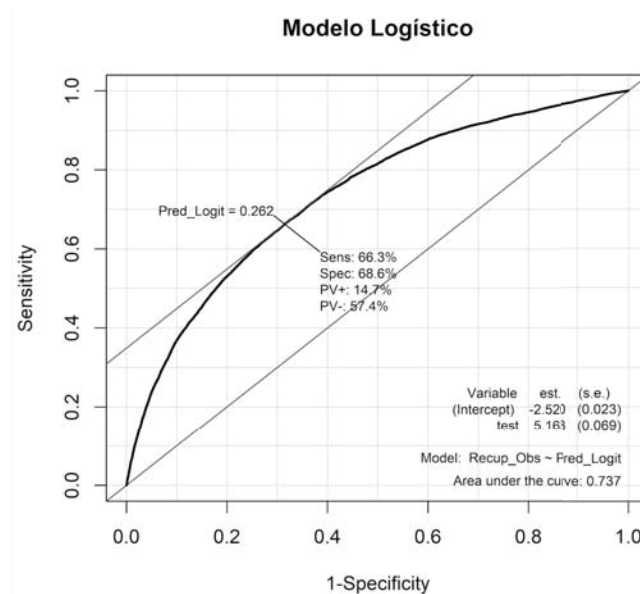
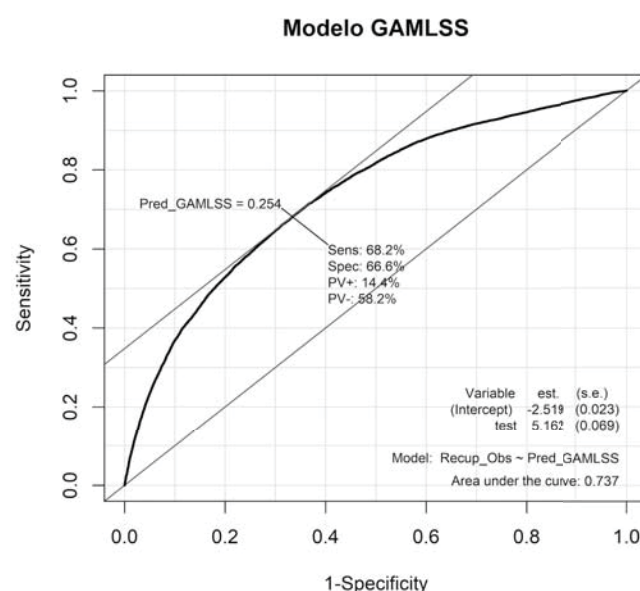
| Covariáveis               |          | Parâmetro |        |        |
|---------------------------|----------|-----------|--------|--------|
|                           |          | $\mu$     | $\nu$  | $\tau$ |
| Intercepto                |          | 1,679     | -2,803 | -3,484 |
| Vlr de Exposição          | V1       | -0,509    |        | -2,078 |
| Máx.Exposição             | V2       | 0,420     | 0,152  | 2,056  |
| Perc.Vencido              | V3       | -0,192    |        |        |
| Perc.Atr.Chq 6M           | I(V4=1)  | 0,241     | 0,406  |        |
| Perc.Atr.Chq 6M           | I(V4=2)  | -0,004    | -0,194 |        |
| Indic.Renegoc             | I(V5=1)  | -0,315    |        |        |
| Restritivo LP A/D         | I(V6=1)  | -1,126    | 1,334  |        |
| Restritivo LP A/D         | I(V6=2)  | -1,525    | 1,884  |        |
| Restritivo LP A/D         | I(V6=3)  | -1,606    | 2,522  |        |
| Perc.Oper.em LP           | V7       | 1,663     | -2,998 |        |
| Perc.Sld.Prejuízo         | V8       | -0,406    |        |        |
| Perc.Sld.Endivid.         | V9       |           | 0,149  |        |
| Qtde Cheque Dev.          | I(V10=1) |           | 0,132  |        |
| Máx.Atraso Reneg.I(V11=1) |          |           | -0,661 |        |
| Máx.Atraso Reneg.I(V11=2) |          |           | -0,502 |        |
| Perc.Máx.6M U.L.          | I(V12=1) |           | -0,312 |        |
| Perc.Máx.6M U.L.          | I(V12=2) |           | -0,435 |        |
| Perc.Máx.12M Atr          | V13      | 0,441     | 0,850  |        |
| Rest.Op.Venc.Atv          | V14      | -1,243    |        |        |
| Rest.Oper.LP Atv          | I(V15=1) | 0,905     | -0,925 |        |
| Rest.Oper.LP Atv          | I(V15=2) | 1,047     | -1,181 |        |
| Rest.Oper.LP Atv          | I(V15=3) | 0,956     | -1,678 |        |
| Rest.Créd.Atv             | V16      | 0,377     |        |        |
| Score de Mercado          | V17      | -0,039    |        |        |
| Máx.dias Atraso           | V18      | 0,156     | 0,235  |        |
| Tempo de Relac.           | V19      | -0,008    |        |        |
| Qtde Restr.014 G4         | V20      | -0,010    |        |        |
| Grau Máx.Rest.Bx          | I(V21=1) | -0,056    |        |        |
| Grau Máx.Rest.Bx          | I(V21=2) | -0,176    |        |        |
| Grau Máx.Rest.Bx          | I(V21=3) | -0,024    |        |        |
| Grau Máx.Rest.Dc          | I(V22=1) | -0,008    |        |        |
| Grau Máx.Rest.Dc          | I(V22=2) | 0,348     |        |        |
| Grau Máx.Rest.Dc          | I(V22=3) | 0,038     |        |        |
| Rest.Op.Venc.Bx           | V23      | -0,100    |        |        |
| Rest.Protest.Bx           | I(V24=1) | -0,141    |        |        |
| Rest.Reneg.Bx             | I(V25=1) | -0,152    |        |        |
| Rest.Reneg.Bx             | I(V25=2) | -0,154    |        |        |
| Perc.CCF Bx               | I(V26=1) | -0,107    |        |        |
| Perc.CCF Bx               | I(V26=2) | -0,259    |        |        |
| Rest.Ref.Atv.3M           | I(V27=1) | -0,170    |        |        |
| Idade Cliente             | V28      | 0,222     |        |        |
| Ind.Renegoc.9M            | I(V29=1) |           | 0,341  |        |
| Score Cred.Ajust.         | V30      |           | -0,118 |        |

Como mencionado no início deste artigo, comumente uma regressão logística é aplicada para para gestão de clientes inadimplentes, sendo assim performou-se um estudo comparativo do modelo gerado GAMLSS e um modelo de regressão logística ajustado sobre os mesmos dados. A Tabela 7 apresenta um erro quadrático médio muito menor para o modelo GAMLSS para todos as safras de validação.

**Tabela 7:** Erro quadrático médio comparativo

| Modelo    | Erro Quadrático Médio |        |        |        |
|-----------|-----------------------|--------|--------|--------|
|           | Ajuste                | Dez/17 | Jan/18 | Fev/18 |
| Logístico | 0,3108                | 0,3169 | 0,2899 | 0,2918 |
| GAMLSS    | 0,1133                | 0,1135 | 0,1142 | 0,1217 |

As Figuras 6 e 7 demonstram as curvas ROC para os dois modelos, uma performance muito similar para os indicadores de sensibilidade, especificidade e área sob a curva.

**Figura 6:** Curva ROC modelo Logístico**Figura 7:** Curva ROC modelo GAMLSS

Finalmente, selecionou-se um pequeno grupo de clientes os quais calculou-se as principais componentes

de saída do modelo, ou seja, a probabilidade de ocorrer nenhum pagamento ( $p_0$ ), a probabilidade do percentual do pagamento estar entre 0,01% e 99,99% ( $p_2$ ), a probabilidade de ocorrer o pagamento total ( $p_1$ ) e a esperança de  $y$  ( $E(y)$ ). Adicionalmente, incluiu-se na Tabela 8 o valor real da variável de interesse ( $Y$ ).

**Tabela 8:** Exemplo de previsões

| Cliente | Nada ( $p_0$ ) | Algo ( $p_2$ ) | Tudo ( $p_1$ ) | $E(y)$ | Real ( $Y$ ) |
|---------|----------------|----------------|----------------|--------|--------------|
| A       | 0.020          | 0.345          | 0.633          | 0.819  | 0.145        |
| B       | 0.979          | 0.018          | 0.001          | 0.006  | 0.000        |
| C       | 0.119          | 0.769          | 0.111          | 0.499  | 0.920        |
| D       | 0.024          | 0.356          | 0.619          | 0.811  | 1.000        |
| E       | 0.033          | 0.802          | 0.164          | 0.548  | 0.464        |
| F       | 0.991          | 0.006          | 0.001          | 0.005  | 0.000        |

## 5. Conclusão

O estudo mostrou-se extremamente válido e viável. Apresentou resultados semelhantes a tradicional método de regressão logística, com a vantagem de fornecer informação relevante para a gestão de portfólio de cobrança bancária, em processos como precificação de carteira e tabela de descontos. A metodologia GAMLSS e a distribuição beta inflacionada em zero e um apresentou resultados satisfatórios e aderentes a uma solução para o problema estabelecido. O modelo gerado com o *framework* GAMLSS demonstrou-se robusto para classificar dados binários tais como a regressão logística, bem como fornece a esperada previsão da taxa de pagamento da dívida.

A partir destes resultados, vislumbrou-se outros estudos que podem ser desenvolvidos tais como para outras carteiras de produtos, bem como, para níveis de atraso mais avançados. Outras distribuições além da  $Beinf(\mu, \sigma, \nu, \tau)$ , podem apresentar melhor aderência aos dados. Também apresenta-se como oportunidade, a exploração de outras funcionalidades que o *framework* GAMLSS pode ofertar.

## Agradecimentos

Primeiramente, agradeço minha esposa Giselle e filhas Alessia e Estella, pelo incondicional apoio a cumprir todas as etapas do curso. Agradeço a coordenação do Curso de *Data Science & Big Data* da UFPR, seu corpo docente e administração que oportunizaram intenso avanço no conhecimento. Ao meu orientador José Luiz Padilha Silva, bem como meu Co-orientador Cesar Augusto Taconeli, pelo incansável suporte e disponibilidade em todos os momentos deste artigo. Aos meus

colegas Bruno Geronymo e Thiago da Silva Alves pelo apoio em diversas fases do período do curso. E finalmente aos meus colegas de curso pelo elevado nível de discussão e pontos de vista, doravante serão sempre referência para mim.

## Referências

- [1] RIGBY R.A., STASINOPOULOS D.M. (2001). *The GAMLSS project: a Flexible Approach to Statistical Modelling*. In B Klein, L Korsholm (eds.), "New Trends in Statistical Modelling: Proceedings of the 16th International Workshop on Statistical Modelling," pp. 249–256. Odense, Denmark.
- [2] AKANTZILIOTOU C., RIGBY R.A. and STASINOPOULOS D.M. (2002). *The R Implementation of Generalized Additive Models for Location, Scale and Shape*. In M Stasinopoulos, G Touloumi (eds.), "Statistical Modelling in Society: Proceedings of the 17th International Workshop on Statistical Modelling," pp. 75–83. Chania, Greece.
- [3] RIGBY R.A., STASINOPOULOS D.M. (2005). *Generalized Additive Models for Location, Scale and Shape*. Applied Statistics, 54, 507–554.
- [4] HASTIE, T. and TIBSHIRANI, R. (1990), *Generalized Additive Models*, Chapman and Hall.
- [5] MANFIO, F. (2003), *O Risco nosso de cada dia.*, São Paulo, Estação das Letras.
- [6] VAN BUUREN, S.; GROOTHUIS-OUDSHOORN, K. (2011), *mice: Multivariate imputation by chained equations in R*. Journal of Statistical Software, v. 45, n. 3, p. 1–67. Disponível em: <https://www.jstatsoft.org/v45/i03/>.
- [7] R CORE TEAM (2018), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, <https://www.R-project.org/>
- [8] SAS Institute Inc (2015), *Base SAS 9.4 Procedures Guide*, <http://support.sas.com>, NC: Sas Institute Inc, 5th ed. Cary
- [9] OSPINA, R. and FERRARI, S.L.P. (2007), *Inflated Beta Distributions*, Statistical Papers, 10.1007/s00362-008-0125-4
- [10] NELDER, J.A. and WEDDERBURN, R.W.M. (1972), *Generalized linear models*, Journal of the Royal Statistical Society: Series A (General), Wiley Online Library