Universidade Federal Fluminense

FACULDADE DE ECONOMIA

PATRICK RIBEIRO MAIA

UMA APLICAÇÃO DA REGRESSÃO LOGÍSTICA E MÉTODOS DE VALIDAÇÃO NA CLASSIFICAÇÃO DE RISCO DE CRÉDITO

NITERÓI – RJ

2019

PATRICK RIBEIRO MAIA

UMA APLICAÇÃO DA REGRESSÃO LOGÍSTICA E MÉTODOS DE VALIDAÇÃO NA CLASSIFICAÇÃO DE RISCO DE CRÉDITO

Monografia apresentada ao curso de Bacharelado em Ciências Econômicas da Universidade Federal Fluminense como requisito parcial para conclusão do curso.

Orientador:
Prof. Dr. Jesus Alexei Luizar Obregon

Niterói – RJ

2019

Ficha catalográfica automática - SDC/BEC
Gerada com informações fornecidas pelo autor

M217a Maia, Patrick Ribeiro
Uma aplicação da regressão logística e métodos de validação na classificação de risco de crédito / Patrick Ribeiro Maia ; Jesus Alexei Luizar Obregon, orientador. Niterói, 2019.
44 f. : il.

Trabalho de Conclusão de Curso (Graduação em Ciências Econômicas)-Universidade Federal Fluminense, Faculdade de Economia, Niterói, 2019.

1. Risco de crédito. 2. Econometria. 3. Regressão logística. 4. Produção intelectual. I. Obregon, Jesus Alexei Luizar, orientador. II. Universidade Federal Fluminense. Faculdade de Economia. III. Título.

CDD -

Bibliotecária responsável: Cláudia Maria Gomes Curi - CRB7/4192

PATRICK RIBEIRO MAIA

UMA APLICAÇÃO DA REGRESSÃO LOGÍSTICA E MÉTODOS DE VALIDAÇÃO NA CLASSIFICAÇÃO DE RISCO DE CRÉDITO

Monografia apresentada ao curso de Bacharelado em Ciências Econômicas da Universidade Federal Fluminense como requisito parcial para conclusão do curso.

Trabalho aprovado em 02 de dezembro de 2019

BANCA EXAMINADORA

Prof. Dr. Jesus Alexei Luizar Obregon
Orientador
Universidade Federal Fluminense

Prof. Dr. André Barbosa Oliveira
Universidade Federal Fluminense

Prof. Dr. Antonio Carlos Fiorencio Soares da Cunha
Universidade Federal Fluminense

RESUMO

O objetivo dessa monografia é estimar um modelo de classificação de risco de crédito utilizando a metodologia convencional e largamente utilizada em instituições financeiras nacionais e internacionais, a regressão logística. Uma vez estimado o modelo de classificação de risco de crédito, procederemos com a validação do mesmo para verificar a capacidade preditiva do modelo e seu impacto no processo de concessão de crédito. Devido a escassez de bases de dados de operações de crédito no mercado nacional, utilizaremos as bases de dados da fintech *Lending Club*, que atua no segmento de microcrédito norte-americano desde o início de 2007 e publica o status de todos os seus empréstimos concedidos desde então, além de inúmeras covariantes socio-econômicas que poderão compor o modelo estimado. Os resultados obtidos comprovam a capacidade da regressão logística de prover um modelo com qualidade satisfatória para ser utilizado por instituições financeiras para mensurar a probabilidade de inadimplência de novos empréstimos.

Palavras-chave: Risco de crédito; Regressão logística; Classificação Binária; Economia.

ABSTRACT

The objective of this work is to estimate a credit risk classification model using the conventional and widely used methodology in national and international financial institutions, the logistic regression. Once the credit risk classification model has been estimated, we will proceed with its validation to verify the predictive capacity of the model and its impact on the credit granting process. Due to the scarcity of databases of credit operations in the domestic market, we will use the Lending Club databases, which operates in the North American microcredit segment since the beginning of 2007 and publishes the status of all its loans granted since then, in addition to numerous socioeconomic covariants that may compose the estimated model. The results obtained prove the capacity of the logistic regression to provide a model with satisfactory quality to be used by financial institutions to measure the probability of default of new loans.

Keywords: Credit risk; Logistic regression; Binary classification; Economics.

“Now I understand, said the last man.”
(Arthur C. Clarke)

Dedico esse trabalho aos meus pais, Regilene e Adriano.

AGRADECIMENTOS

Agradeço primeiramente aos meus pais, que ofereceram o apoio viabilizador de absolutamente toda e qualquer conquista minha ao longo da vida.

Agradeço aos excelentes professores que se dedicaram em compartilhar mais que o conhecimento técnico, mas a verdadeira motivação que há por trás do estudo das ciências econômicas. Necessário oferecer agradecimentos mais que especiais para os professores Jesus Alexei Luizar Obregon pelas excelentes aulas de econometria e métodos quantitativos e Renault Michael pelas densas aulas de história macroeconômica brasileira.

Agradeço imensamente a Vitória por todo o carinho e apoio que me proporciona e ainda proporcionará, por sempre me lembrar que eventualmente tudo vai dar certo e pelo enorme presente que me dá diariamente ao me permitir conhecer a pessoa maravilhosa que você é.

Agradeço a todo o corpo funcional do BNDES com quem tive contato ao longo da minha primeira experiência profissional como economista, em especial a Marlon Tyrone a quem atribuo a maior parte de todo o meu conhecimento e interesse na área de risco de crédito.

Agradeço à todos os idealizadores da Mutual que desde sempre enxergaram e enxergam o meu potencial, tenho certeza que alcançaremos juntos o objetivo de tornar mais justo o mercado de crédito brasileiro. Agradeço a Pedro Cavalcante e Eduarda Oliveira, os cientistas de dados mais promissores do Brasil e que diariamente me surpreendem com a sua capacidade de inovar e de resolver problemas abstratos com muita criatividade.

Ao meu grande amigo ao longo de todos os semestres da graduação, um obrigado à Carlos Salim pela honra de todas as conversas, conselhos e revisões de última hora.

Agradeço a Giovana por todas as palavras de calma em meio ao caos do ensino superior e por ser a grande amiga você diz que eu não mereço ter.

Ao mestre e futuro doutor Daniel Duque, obrigado por toda a confiança e todos os excelentes e bem sucedidos projetos que desenvolvemos e ainda vamos desenvolver juntos.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

É conhecido que no Brasil, há 2 anos a taxa básica de juros (SELIC) retornou ao patamar de 1 dígito, porém, as taxas de juros cobradas nas modalidades mais abrangentes de crédito (cartão de crédito, cheque especial e empréstimo pessoal sem garantia) podem alcançar 300% a.a. Essa distância entre a taxa livre de risco e a taxa efetivamente cobrada da população é muitas vezes atribuídas ao alto risco de se emprestar para pessoas físicas no Brasil devido a inadimplência.

A correta estimação do risco de inadimplência no processo de concessão de crédito de varejo é de extrema importância dado que tanto a precificação quanto a saúde da instituição financeira dependem desse parâmetro. Uma vez que as instituições financeiras de grande porte precisam avaliar o risco de crédito de milhares de operações diariamente, o processo de concessão evoluiu para que a intervenção humana no processo decisório fosse cada vez menor ao longo do tempo, de forma que modelos econométricos tomaram o lugar da figura dos analistas de risco de crédito.

Modelos estatísticos para previsão de risco de crédito estão sendo desenvolvidos desde Greene 1992 e nos mais de 27 anos do aperfeiçoamento desse segmento, muitas técnicas foram incorporadas, porém, a tradicional regressão logística ainda apresenta desempenho competitivo em relação à técnicas mais inovadoras como as que envolvem redes neurais e *random forests*.

O objetivo dessa monografia é contribuir com o desenvolvimento e implementação das técnicas estatísticas, inclusive no que diz respeito a aplicação computacional, no segmento do risco de crédito, uma vez que o uso de um ferramental sofisticado na resolução dessa classe de problemas pode contribuir com a formação de um mercado de crédito mais justo e barato para todos os participantes que se comprometem a honrar os termos contratuais da operação de empréstimo.

Nesse estudo, iniciamos no capítulo 2 com uma revisão de literatura dos conceitos de regressão logística, o arcabouço necessário para estimação e validação de modelos de classificação binária, no capítulo 3 expomos a implementação e estimação do modelo de classificação de risco de crédito e seus coeficientes e no capítulo 4 realizamos a validação da performance do modelo estimado. Finalmente apresentamos as considerações finais, junto com os códigos computacionais e outras informações adicionais nos respectivos apêndices.

2 REVISÃO DE LITERATURA

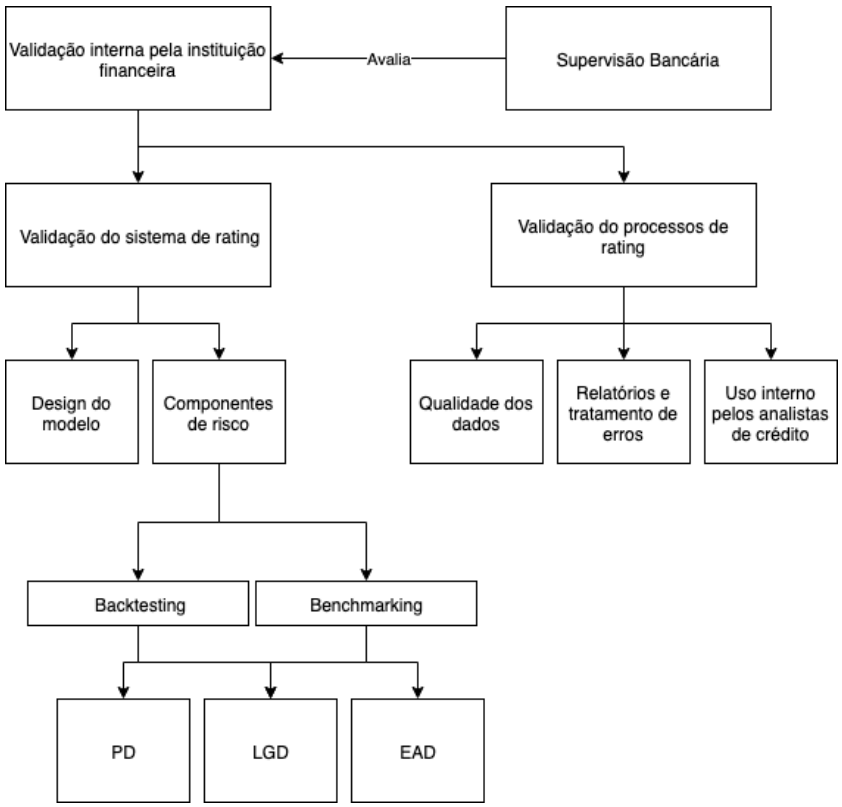
A metodologia de estimação do modelo de classificação de risco de crédito desenvolvido nesse trabalho se baseia nos trabalhos de Greene 1992 no qual é apresentado o potencial da regressão logística para estimar a probabilidade de inadimplência a partir de variáveis socio-econômicas e financeiras. No seu trabalho seminal, o autor expõe um esquema para o uso da regressão logística nesse segmento:

- Aquisição de dados históricos contendo as potenciais variáveis a serem incluídas no modelo.
- Marcação da variável independente (inadimplência) a partir de um critério absoluto e não variante no tempo.
- Estimação do modelo através de regressão logística prezando pela parcimônia (menos variáveis).

A validação e a avaliação de performance dos modelos de classificação binária possuem ainda mais importância quando o objetivo é a classificação de risco de crédito pois todo o processo possui princípios definidos pelo Comitê de Supervisão Bancária de Basileia que devem nortear o processo de desenvolvimento, validação e acompanhamento dessa classe de modelos. Um resumo desse processo é descrito na Tabela 1, de acordo com os normativos do BCBS (Basel Committee on Banking Supervision) como publicado em Committee et al. 2005.

A entidade de supervisão bancária, no Brasil representada pelo Banco Central do Brasil, avalia o arcabouço de validação interna pela instituição financeira, contemplando o design do modelo e os componentes discriminantes de risco utilizados em sua formulação. Essa validação contempla a qualidade dos dados analisados e determina a exigência de relatórios periódicos sobre sua performance e possíveis erros decorrentes do seu uso em produção. No que diz respeito à validação prática, a instituição financeira deve fazer uso de *backtesting* e *benchmarking* dos seus modelos, ou seja, avaliar o modelo atual em dados históricos e ter métricas de performance que são comparáveis no tempo a fim de avaliar a variação do desempenho do modelo ao longo da concessão de empréstimos. O *backtesting* e o *benchmarking* deve ser realizado, no mínimo, nas 3 variáveis fundamentais do risco de crédito: a PD (Probabilidade de Inadimplência), a LGD (Perda Dado o Default, do inglês, *Loss Given Default*) e a EAD (Exposição no Default, do inglês, *Exposure at Default*).

Figura 1 – Metodologia de Desenvolvimento - BCBS.



Fonte: Elaboração própria.

2.1 Regressão Logística

Nesta seção, abordaremos os conceitos relativos à representação econométrica do modelo de regressão logística, alertamos que o conteúdo aqui apresentado pode ser encontrado nas referências Nelder e Wedderburn 1972 e Gourieroux 2000. A regressão logística participa da família de modelos lineares generalizados e é definida a partir de uma variável resposta dicotômica que corresponde a realização ou não do evento analisado. Essa variável assume o valor 1 caso seja observado evento de inadimplência dentro da janela de acontecimentos e 0 caso o empréstimo continue adimplente ao longo de toda a vida útil da operação:

$$y_i = \begin{cases} 1, & \text{se inadimplente} \\ 0, & \text{se adimplente} \end{cases} \tag{2.1}$$

A probabilidade aqui modelada é uma função não linear de uma combinação linear de fatores potencialmente discriminantes do risco da operação, definiremos esses fatores como sendo iguais ao vetor de variáveis endógenas $x_i^{'}$ ao longo da exposição da metodologia de estimação.

Resta definir a relação entre a variável dependente e as variáveis socioeconômicas, a forma usual é representada por um modelo linear em função de variáveis explicativas que podem

ser condensadas em uma única variável (x_i) para fins de desenvolvimento do modelo mas que pode ser estendido para conter mais variáveis:

$$yi^* = x_i\beta + \mu_i, \quad i = 1, ..., n \quad , \tag{2.2}$$

onde yi^* não é diretamente observada e é convencionalmente chamada de variável "latente". O que observamos é uma variável binária definida por:

$$y_i = \begin{cases} 1, & \text{se } y^* > 0 \\ 0, & \text{caso contrário} \end{cases} .$$

A partir das equações (2.1) e (2.2) podemos definir a distribuição de y e o problema de regressão logística:

$$\begin{aligned} P(y_i = 1|X_i = x_i) &= P(y_i^* < \ell_i) \\ &= P(x_i\beta + \mu_i < \ell_i) \\ &= P\left(\frac{\mu_i}{\sigma} < \frac{\ell_i}{\sigma} - \frac{x_i\beta}{\sigma}\right) \\ &= F\left(\frac{\ell_i}{\sigma} - \frac{x_i\beta}{\sigma}\right) \\ &= P_i \text{ (por construção)} \quad , \end{aligned} \tag{2.3}$$

onde ℓ_i representa o ponto de corte que separa as observações observadas como inadimplentes ou como adimplentes. Devemos ressaltar que como a equação (2.3) não depende da mudança de escala, ou seja, das variações de σ, podemos considerar $\sigma = 1$ para continuar o desenvolvimento do modelo.

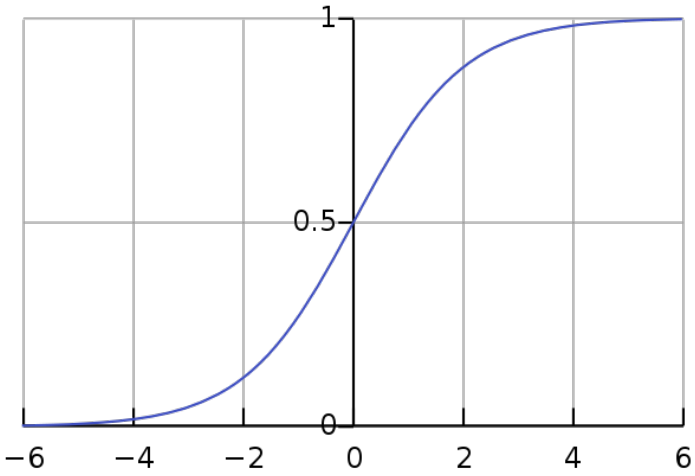
Os parâmetros de risco que influenciam a probabilidade de inadimplência (PD) são expressos pelo vetor de parâmetros β. Esses parâmetros podem ser obtidos pelo método de máxima verossimilhança que conduzem a problemas não lineares que podem ser abordados pelas clássicas técnicas numéricas ou ferramentas mais sofisticadas como mínimos quadrados iterativamente ponderados. (*IWLS*).

A regressão logística depende da definição de função logística para ser estimada, portanto, definiremos aqui a função logística como uma sigmoide que recebe um valor real x ($x \in \mathbb{R}$) e retorna um valor entre 0 e 1. A função logística padrão é definida na equação (2.4):

$$F(x) = \frac{e^x}{e^x + 1} = \frac{1}{1 + e^{-x}} \quad . \tag{2.4}$$

Na Figura 2 representamos graficamente a função logística e vemos que quando $x \to \infty$, a função tende a 1 e quando $x \to -\infty$, a função tende a 0.

Figura 2 – Função logística



Fonte: Elaboração própria.

2.1.1 Estimação de Máxima Verossimilhança

Podemos recorrer a Gourieroux 2000 para uma exposição completa do processo de estimação no contexto de máxima verossimilhança. Como é conhecido, a função de máxima verossimilhança é definida:

$$L(y;b) = \prod_{i=1}^{n} \{F(x_i b)^{y_i}[1 - F(x_i b)]^{1-y_i}\} \quad , \tag{2.5}$$

onde b representa o parâmetro que otimiza a função de máxima verossimilhança.

Seguimos a partir da equação de máxima verossimilhança (2.5) o procedimento tradicional e partir dele podemos obter a função de log-verossimlhança:

$$\begin{aligned} log(L) &= \sum_{i=1}^{n} y_i log[F(x_i b)] + (1 - y_i)[1 - F(x_i b)] \\ log(L) &= \sum_{i:y_i=1}^{n} log[F(x_i b)] + \sum_{i:y_i=0}^{n} log[1 - F(x_i b)] \quad . \end{aligned} \tag{2.6}$$

Uma condição suficiente para afirmar que existe um único máximo global de $log(L)$ se baseia no fato de que essa função é estritamente concava, ou, de forma equivalente, que $log(F)$ e $log(1 - F)$ são estritamente côncavos e essa condição se sustenta para o modelo de regressão logística, e com essa propriedade podemos nos apoiar nas ferramentas de cálculo diferencial, para derivar a equação (2.6) com respeito ao vetor de parâmetros b e assim obter o vetor de

derivadas igual a 0 nos pontos de máximo a partir das condições de primeira ordem:

$$\frac{\partial logL}{\partial b} = \sum_{i:y_i=1} \frac{f(x_i b)}{F(x_i b)} x_i^{'} - \sum_{i:y_i=0} \frac{f(x_i)}{1 - F(x_i b)} x_i^{'} = 0 \quad , \tag{2.7}$$

onde f é a função de densidade associada a F e $x_i^{'}$ representa a transposição do vetor contendo a variável explicativa.

$$0 = \sum_{i=1}^{n} \left[\frac{y_i}{F(x_i b)} - \frac{1 - y_i}{1 - F(x_i b)} \right] f(x_i b) x_i^{'} = \sum_{i=1}^{n} \frac{y_i - F(x_i b)}{F(x_i b)[1 - F(x_b)]} f(x_i b) x_i^{'} \quad . \tag{2.8}$$

Podemos simplificar a equação (2.8) uma vez que a distribuição de densidade da sigmoide associada a função logística pode ser escrita:

$$f(x) = F(x)[1 - F(x)] = \frac{e^{-x}}{(1 + e^{-x})^2} \quad , \tag{2.9}$$

aplicando a equação (2.9) na equação (2.8), obtemos:

$$0 = \sum_{i=1}^{n} [y_i - F(x_i b)] x_i^{'}$$

$$\sum_{i=1}^{n} y_i x_i^{'} = \sum_{i=1}^{n} F(x_i b) x_i^{'} \quad . \tag{2.10}$$

Gourieroux 2000 ressalta que as as equações de verossimilhança associadas ao modelo de regressão logística não possuem parâmetros lineares, portanto, precisamos empregar o uso de soluções numéricas, tais como o método de Newton Raphson e o método de Scoring.

2.1.2 Newton Raphson e Scoring

De uma forma geral e simplificada o sistema de equações não lineares $G(z) = 0$ é resolvido segundo o método de Newton admitindo que G é diferenciável e é válida a aproximação:

$$G(z) \approx G(z_0) + G^{'}(z_0)(z - z_0) \quad ,$$

onde z_0 é dado e $G^{'}$ é uma transformação linear definida em espaços apropriados. Considerando que z deve ser uma raiz de G então:

$$-G(z_0) = G^{'}(z_0)(z - z_0) \quad ,$$

admitindo que $G^{'}$ possui inversa ($G^{'-1}$):

$$
\begin{aligned}
G^{'}(z_0)(z - z_0) &= -G(z_0) \\
(z - z_0) &= (G^{'}(z_0))^{-1}(G(z_0) \\
z &= z_0 - (G^{'}(z_0))^{-1}G(z_0) \quad .
\end{aligned} \tag{2.11}
$$

No problema aqui analisado, o método de Newton Raphson é aplicado à regressão logística, com o objetivo de encontrar a raiz da equação (2.7):

$$G = \frac{\partial log(L)}{\partial b} = 0$$

e possui a seguinte equação objetivo a ser otimizada:

$$\beta_{h+1} = \beta_h - \left\{ \frac{\partial^2 log[L(\mathbf{b})]}{\partial \beta_h \partial \beta_h'} \right\}^{-1} \frac{\partial log[L(\mathbf{b})]}{\partial \beta_h} \tag{2.12}$$

Segundo Gourieroux 2000, o método de scoring difere do método de newton ao passo que envolve a substituição de parte da equação objetivo por uma expectativa condicional, ou seja, a substituição do termo:

$$\frac{\partial^2 log[L(\mathbf{b})]}{\partial \beta \partial \beta'}$$

na equação (2.12) pela expectativa condicional de x, gerando uma nova função objetivo:

$$\beta_{h+1} = \beta_h + E \left\{ -\frac{\partial^2 log[L(\mathbf{b})]}{\partial \beta_h \partial \beta_h'} \right\}^{-1} \frac{\partial log[L(\mathbf{b})]}{\partial \beta_h}$$

2.1.3 IWLS

O desenvolvimento do algoritmo proposto por Dutang 2017 tem como base a teoria de modelos lineares generalizados no qual são definidos conceitos da família exponencial, funções de enlace e outros conceitos. E demonstrado que a regressão logística pode ser considerado como um caso especial de uma função que pertence à família exponencial com certas correspondências a Dobson e Barnett 2008. Assim a nomenclatura utilizada por Dutang corresponde às desenvolvidas na literatura de modelos lineares generalizados (GLM). Considerando que esse tema vai além do objetivo desta monografia decidimos somente apresentar o algoritmo de mínimos quadrados iterativamente ponderados (IWLS) a seguir:

Algoritmo 2.1.1: IWLS

Entrada: y_i e X

Saída: Parâmetros β

Início: IWLS

`/* Inicialização do algoritmo */`

1 Usar os dados originais com um pequeno desvio $\mu^{(0)} = y_i + 0.1$ para calcular $\eta_i^{(0)} = g(\mu_i^{(0)})$

2 Calcular o conjunto de respostas $Z^{(0)} = (\eta^{(0)} + (y_i - \mu^{(0)})(g^{'}(\mu^{(0)})_i)$

3 Calcular o conjunto inicial de pesos

$$W^{(0)} = diag(w_1, ..., w_n)$$
$$w_i = \frac{1}{a(\phi_i)(g^{'}(\mu_i^{(0)}))^2 V(\mu_i^{(0)})}$$

4 Resolver o sistema para obter $\beta^{(0)}$:

$$X^T W^{(0)} X \beta^{(0)} = X^T W^{(0)} Z^{(0)}$$

`/* Uma vez inicializado o algoritmo, podemos realizar as iterações necessárias para a estimação. */`

5 **Para** $k = 1, ..., m$ **faça**

6 Calcular o conjunto de respostas

$$Z^{(k)} = (z_i)_i$$
$$z_i = \eta_i(\beta^{(k)}) + (y_i - \mu_i(\beta^{(k)}))g^{'}(\mu_i \beta^{(k)}))$$

7 Calcular o conjunto de pesos atualizado

$$W^{(k)} = diag(w_1, ..., w_n)$$
$$w_i = \frac{1}{a(\phi_i)(g^{'}(\mu_i^{(\beta^{(k)})}))^2 V(\mu_i^{(\beta(k))})}$$

8 Resolver o sistema para obter $\beta^{(k+1)}$:

$$X^T W^{(k)} X \beta^{(k+1)} = X^T W^{(k)} Z^{(k)}$$

9 Verificar convergência:

$$||Dev(\beta^{(k+1)} - \beta^{(k)})|| \leq \epsilon$$

Fim: IWLS

No algoritmo 2.1.1, a, b, c e d são funções suavizadoras dos parâmetros, V é uma função de proporção da variância e g é a função de enlace da regressão logística, correspondente a equação (2.4).

2.1.4 Aplicação do Modelo

Depois de resolver o sistema obtido a partir de $\frac{l(\beta;x_i)}{\beta} = 0$, obtemos o vetor de estimativas de β. Esse vetor representa uma das muitas vantagens da regressão logística para problemas de classificação binária pois são diretamente interpretáveis como sendo as probabilidades de ocorrência do evento modelado.

Se β é o vetor de parâmetros estimado e x_i é o vetor de variáveis endógenas explicativas do modelo:

$$p(x) = \frac{1}{1 + e^{-\beta x}} = \frac{e^{\beta x}}{1 + e^{\beta x}}$$

$$\frac{p(x)}{1 - p(x)} = e^{\beta x} \quad . \tag{2.13}$$

A equação (2.13) expressa a chance, ou seja, a probabilidade de ocorrência de um evento dividida pela probabilidade de não ocorrência do mesmo evento, dessa observação vir a ser observada como inadimplente ($y_i = 1$).

2.2 Metodologia de validação

Uma vez estimado um modelo de classificação binária, precisamos definir um conjunto de métricas que avaliam a performance e o poder preditivo desse modelo. Essas métricas de performance servem para ao mesmo tempo identificar potenciais problemas na estimação (caso a performance seja muito baixa ou muito alta) e também serve como uma prévia do desempenho desse modelo uma vez colocado em produção para classificar o risco de crédito das operações. O mercado e a academia concentram os esforços da validação em 2 métricas principais, compartilhados com uma classe de modelos mais sofisticados que envolvem Machine Learning: AUROC e KS.

2.2.1 Curva Característica de Operação do Receptor (ROC)

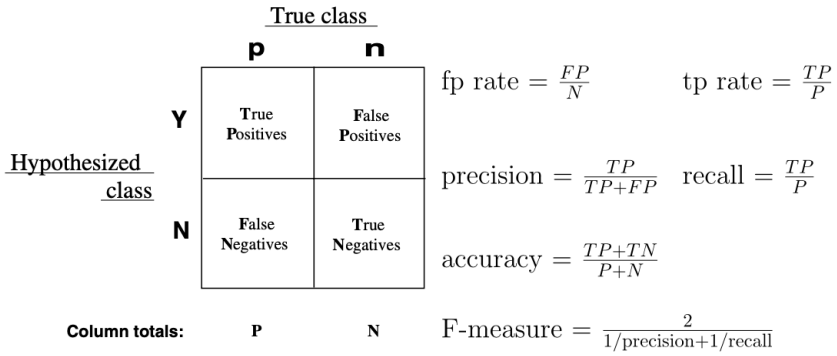
A curva de recepção do operador, do inglês *Receiver Operating Characteristic Curve* (ROC) é uma técnica visual de avaliação de performance de um classificador binário, vamos seguir a exposição realizada por Fawcett 2006 para alcançar a formulação da curva ROC e posteriormente da área formada abaixo dessa curva. Formalmente, um classificador é uma função que mapeia a saída contínua de um modelo com um evento diretamente observável pertencente a um dos elementos do conjunto $\{p, n\}$ (positivo ou negativo para a ocorrência do evento). A partir

da saída contínua de um modelo, podemos estabelecer um ponto de corte *cutoff* que delimita a classificação como positiva ou negativa. Para distinguir o resultado do classificador com a ocorrência do evento, vamos usar o conjunto $\{Y, N\}$ para representar o resultado gerado pelo classificador.

Matriz de Confusão (Confusion Matrix)

Com os conjuntos de observações do evento e de resultados gerados pelo classificador, existem 4 cenários possíveis: se o evento foi observado como positivo e classificado como positivo, temos um caso de positivo verdadeiro (*true positive*); se o evento foi observado como positivo e classificado como negativo, temos um caso de falso negativo (*false negative*); se o evento foi observado como negativo e classificado como negativo, temos um caso de negativo verdadeiro (*true negative*); se o evento foi observado como negativo e classificado como positivo, temos um caso de falso positivo (*false positve*). A partir do resultado de um classificador e das observações dos eventos, podemos formar uma matriz de dupla entrada chamada de matriz de confusão (*confusion matrix*) que fundamenta a formação da curva ROC conforme exposto na Figura 3. As duas principais métricas, que compõem a curva ROC são a taxa de falsos positivos e a taxa de positivos verdadeiros:

Figura 3 – Matriz de Confusão e métricas obtidas a partir dela



Fonte: Fawcett 2006

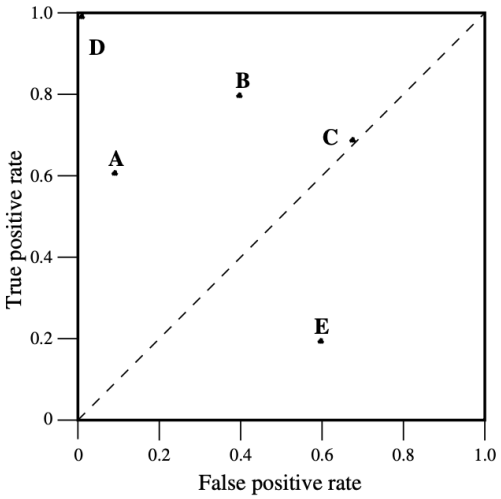
$$\text{tp rate} \approx \frac{\text{Positivos corretamente classificados}}{\text{Positivos totais}}$$

$$\text{fp rate} \approx \frac{\text{Negativos incorretamente classificados}}{\text{Negativos totais}}$$

As curvas ROC são expostas em um gráfico bidimensional no qual a taxa de positivos verdadeiros está disposta no eixo vertical enquanto que a taxa de falsos positivos está disposta no eixo horizontal, representando assim o *tradeoff* entre os benefícios (positivos corretamente

classificados) e os prejuízos (positivos incorretamente classificados). A Figura 4 exibe o resultado de 5 classificadores rotulados entre A e E, e alguns pontos desse gráfico são muito importantes para uma completa compreensão da curva a ser calculada. O ponto à esquerda inferior (0,0) representa um classificador que nunca emite uma classificação positiva, de forma que não há nenhum caso de falso positivo mas também não há nenhum positivo verdadeiro. O classificador perfeito se localiza no ponto (1,0) no qual não há nenhum caso de falso positivo e todas as observações positivas são corretamente classificadas como tal. De forma prática, um classificador é melhor que outro se esse se localiza mais acima e à esquerda, isso é, com uma taxa de positivos verdadeiros mais elevada e uma taxa de falso positivos menor.

Figura 4 – Uma curva básica contendo o resultado de 5 classificadores



Fonte: Fawcett 2006

A linha diagonal ($y = x$) disposta na Figura 4 representa uma performance aleatória, isso é, a estratégia de classificar aleatoriamente uma observação. Por exemplo, se 50% das observações de um evento forem classificadas como positivas e 50% das demais como negativas, é esperado que esse classificador aleatório se localize no ponto (0.5, 0.5), logo, qualquer classificador que se localiza abaixo da diagonal está performando pior que um classificador aleatório.

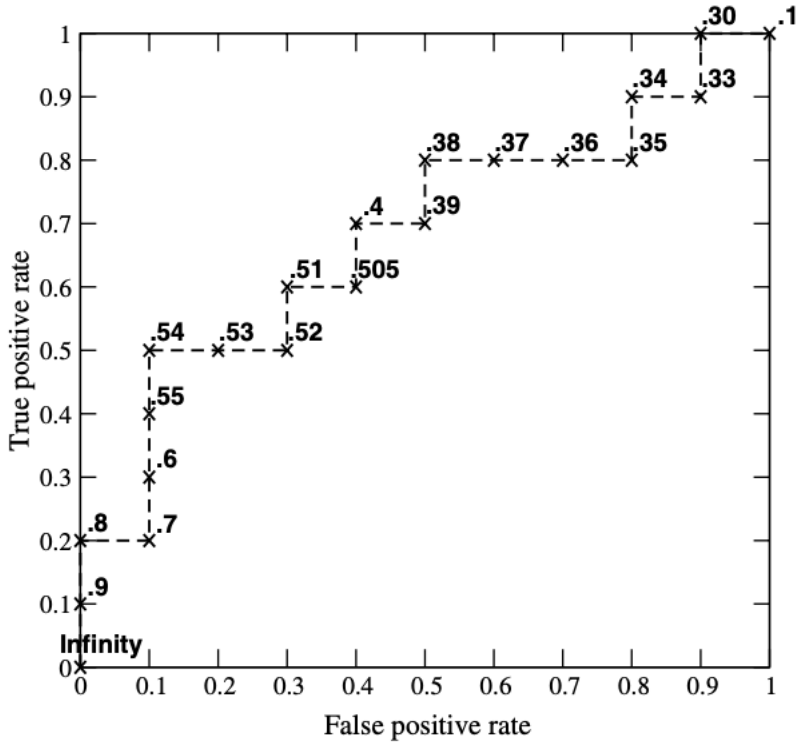
Podemos exemplificar o processo a partir dos dados dispostos na Tabela 1 que contém o resultado de 20 observações avaliadas por um classificador binário, os pontos então compõem o gráfico disponível na Figura 5 no qual é exibido a curva ROC correspondente com cada ponto representando o par gerado a partir de um ponto de corte.

Tabela 1 – Tabela com 20 pontos classificados e o seu score correspondente.

# Obs	Classe	Score	# Obs	Classe	Score
1	p	0.90	11	p	0.40
2	p	0.80	12	n	0.39
3	n	0.70	13	p	0.38
4	p	0.60	14	n	0.37
5	p	0.55	15	n	0.36
6	p	0.54	16	n	0.35
7	n	0.53	17	p	0.34
8	n	0.52	18	n	0.33
9	p	0.51	19	p	0.30
10	n	0.505	20	n	0.10

Fonte: Fawcett 2006

Figura 5 – Curva ROC gerada pelos dados da Tabela 1



Fonte: Fawcett 2006

Uma vez obtida uma função de classificação binária, devemos proceder à geração eficiente da curva ROC que por sua vez explora a monotonicidade das classificações geradas, dado que para qualquer instância classificada como positivo em respeito a um ponto de corte, essa mesma instância também será classificada como positiva para pontos de cortes inferiores, portanto, podemos calcular a curva ROC com um algoritmo eficiente de escaneamento linear como proposto por Fawcett 2006 no Algoritmo 2.2.1.

Algoritmo 2.2.1: ROC

Entrada: L, o conjunto de exemplos de teste; $f(i)$, a probabilidade atribuída pelo estimador que o exemplo i é positivo (inadimplente); P e N, o número de exemplos positivos e negativos

Saída: R, uma lista com os pontos de ROC aumentando de acordo com a taxa de falsos positivos

Requer : $P > 0$ e $N > 0$

Início: ROC

1 $L_{ordenado} \leftarrow L$ ordenado por f de forma decrescente
2 $FP \leftarrow TP \leftarrow 0$
3 $R \leftarrow \langle\rangle$ `/* Vetor Vazio */`
4 $f_{prev} \leftarrow -\infty$
5 $i \leftarrow 1$
6 **Enquanto** $i \leq |L_{ordenado}|$ **faça**
7 **Se** $f(i) \neq f_{prev}$ **então**
8 adiciona $\left(\frac{FP}{N}, \frac{TP}{P}\right)$ no conjunto$R$
9 $f_{prev} \leftarrow f(i)$
10 **Se** $L_{ordenado}[i]$ *é um exemplo positivo* **então**
11 $TP \leftarrow TP + 1$
12 **Senão**
13 $FP \leftarrow FP + 1$
14 $i \leftarrow i + 1$
15 adiciona $\left(\frac{FP}{N}, \frac{TP}{P}\right)$ no conjunto$R$

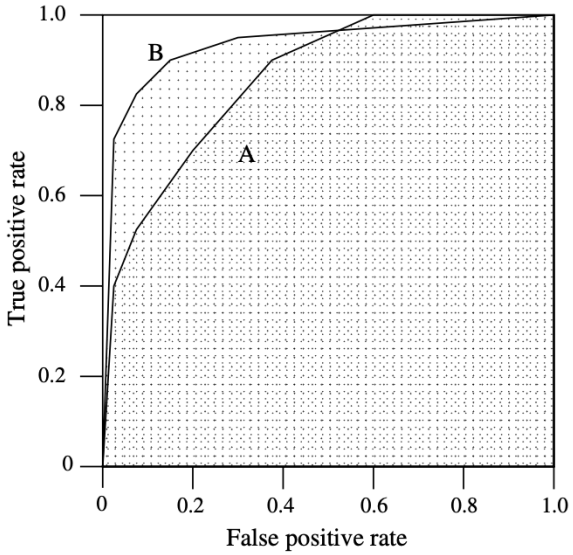
Fim: ROC

Para obter um escalar que representa a performance do classificador binário com base na sua curva ROC, podemos calcular a área abaixo da curva ROC, ou seja, a *Area Under ROC* ou AUROC. Sendo $ROC(y_i, f(i))$ uma função que retorna a curva ROC a partir do resultado de um classificador binário e a variável dependente dicotômica, podemos expressar a AUROC como a integral dessa função, uma vez que quando $n \to \infty$ temos o caso contínuo conforme exposto na Figura 6:

$$AUROC = \int_0^1 ROC(y_i, f(i))$$

No caso discreto, como o da Figura 5, a AUROC pode ser calculada como a área a partir da soma dos retângulos gerados abaixo dos pontos disponíveis.

Figura 6 – AUROC.



Fonte: Fawcett 2006

O desempenho do modelo pode ser avaliado pela AUROC e Oliveira e Andrade 2002 define patamares para o grau de ajustamento de modelos de classificação de risco de crédito a partir dessa métrica:

Tabela 2 – Capacidade de discriminação - AUROC

AUROC	Capacidade de discriminação
0.5	Não existe discriminação
0.7 <= AUROC <= 0.8	Discriminação aceitável
0.8 <= AUROC <0.9	Discriminação excelente
AUROC >= 0.9	Discriminação acima do comum

Fonte: Oliveira e Andrade 2002

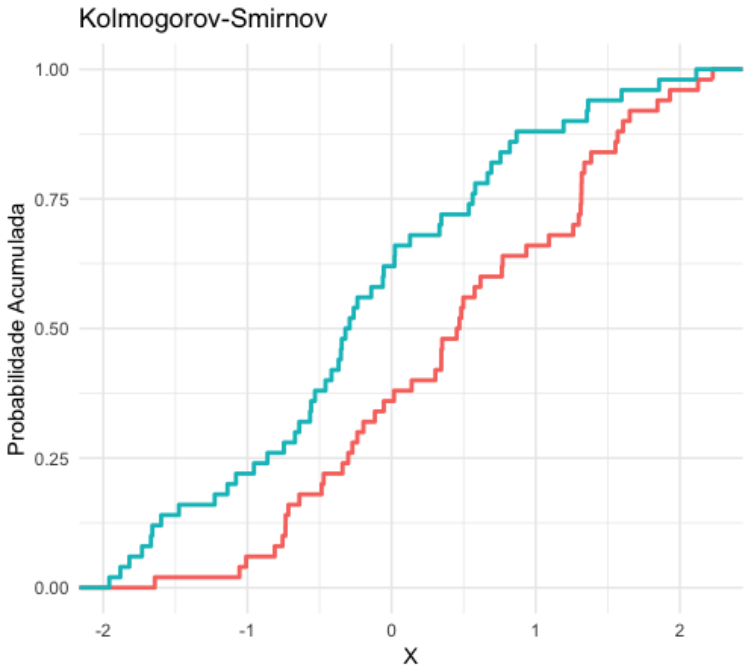
Uma AUROC muito baixa, onde não há discriminação, expõe um problema de insuficiência de variáveis com forte poder preditivo ou um problema na especificação da marcação de inadimplência enquanto que uma AUROC muito alta indica que o modelo apresenta *overfit*.

2.2.2 Kolmogorov-Smirnov

A estatística Kolmogorov-Smirnov (KS) mede a capacidade do modelo de discriminar bons e maus clientes através de um teste não paramétrico formado pela diferença máxima entre as distribuições acumuladas de bons e maus clientes gerada pelo modelo de classificação binária.

De acordo com Arnold e Emerson 2011, se $F_{1,n}(x)$ é a distribuição acumulada das observações de maus tomadores e $F_{2,m}(x)$ é a distribuição acumulada das observações de bons

Figura 7 – Kolmogorov-Smirnov.



Fonte: Elaboração própria.

tomadores, a estatística de Kolmogorov-Smirnov é:

$$D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)| \tag{2.14}$$

A estatística KS é avaliada de acordo com a tabela de qualidade de ajustamento proposta por Oliveira e Andrade 2002 na Tabela 3:

Tabela 3 – Capacidade de discriminação - KS

Valores de KS	Nível de discriminação
Abaixo de 20%	Baixa discriminação
De 20% a 30%	Discriminação aceitável
De 30% a 40%	Boa discriminação
De 40% a 50%	Excelente discriminação
Acima de 50%	Não são muito comuns

Fonte: Oliveira e Andrade 2002

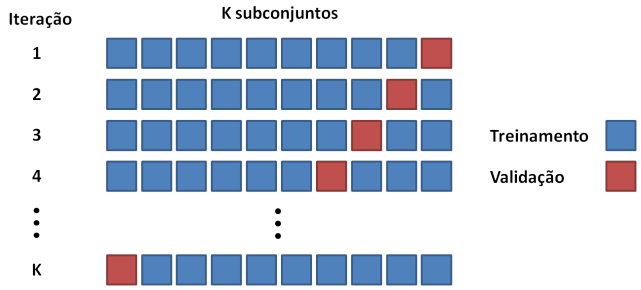
2.2.3 Validação Cruzada

Um modelo de classificação binária pode sofrer viés amostral caso uma característica/-variável muito discriminante esteja concentrada em uma determinada proporção da amostra

de treinamento, isso pode levar a uma performance sub-ótima do modelo uma vez posto em produção.

Podemos diagnosticar esse problema de viés amostral realizando uma validação utilizando o método *K-fold Cross Validation* conforme exposto por Kohavi et al. 1995, através desse método, o conjunto de treinamento é dividido em K subconjuntos de tamanho igual, sendo que apenas K-1 serão utilizados para o treinamento do modelo e 1 subconjunto é utilizado para validação. Esse processo é repetido K vezes de forma que em cada repetição os subconjuntos de treinamento e validação são distintos entre si, ou seja, uma *Cross Validation* com 5-folds terá esse processo repetido 5 vezes, o diagrama da figura 8 expõe o processo.

Figura 8 – K-Fold Cross Validation.



Fonte: Elaboração própria.

Caso a performance entre cada iteração seja substancialmente diferente das demais, então há evidências para a existência de viés amostral e o analista deve retornar à etapa de preparação de dados e análise exploratória para eliminar potenciais variáveis problemáticas.

3 APLICAÇÃO DO MODELO

Na ausência de uma base de dados nacional, faremos uso da base de dados pública oferecida pela fintech norte-americana Lending Club. No seu portal de estatísticas Lending Club Statistics 2019, a fintech oferece a opção de realizar o download de todos os empréstimos concedidos ou reprovados na análise de crédito desde o início das operações em 2007. Cada conjunto de dados é acompanhado de um dicionário bem especificado embora existam alguns problemas de inconsistência na disponibilidade de algumas informações devido ao fato de entrada de novas variáveis na esteira de concessão de crédito. Neste trabalho, analisaremos os dados de empréstimos concedidos pela fintech entre **janeiro de 2017 e dezembro de 2018**. Todos os tratamentos, estimações e visualizações dessa sessão farão uso do software R, desenvolvido por R Core Team 2018.

3.1 Análise Exploratória

O arquivo contendo os empréstimos concedidos entre janeiro de 2017 e dezembro de 2018 contém inicialmente 938821 observações separados em 152 colunas, porém, apenas 49 foram selecionadas para participar da análise exploratória e posteriormente da estimação do modelo, o Anexo 1 contém o nome das variáveis e uma sucinta descrição de seu conteúdo.

3.2 Variável Resposta

Precisamos definir a variável resposta que será utilizada em breve na estimação do modelo, para isso podemos recorrer à variável *loan_status*, cuja distribuição está exposta na Figura 9.

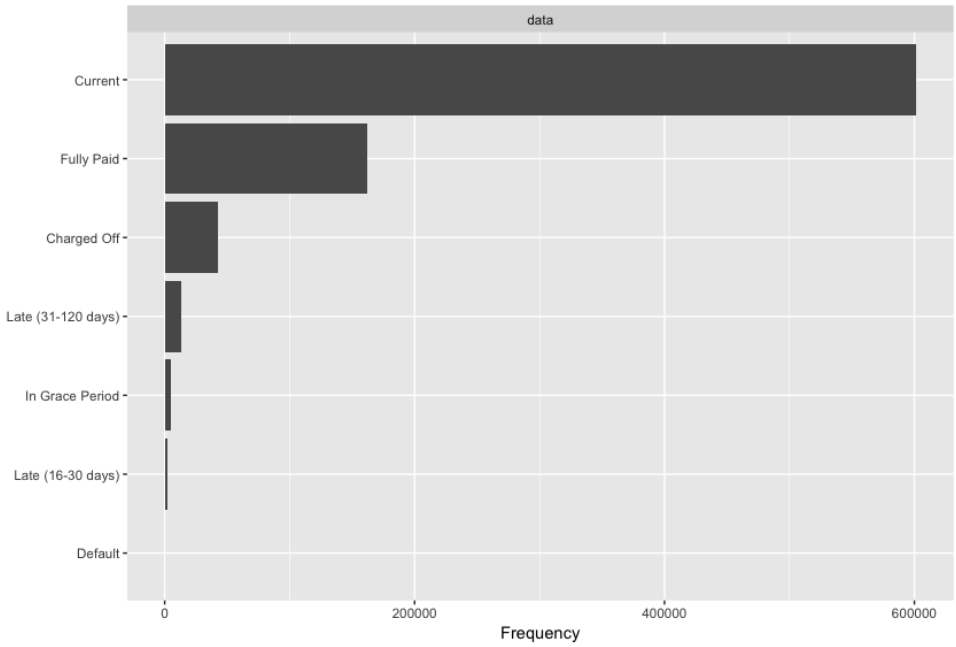
Tabela 4 – Descrição - Status do Pagamento

Loan Status	Descrição
Current	Adimplente e em andamento
Fully Paid	Completamente pago
Charged Off	Em cobrança externa
Late (31 - 120 days)	Atraso entre 31 e 120 dias
In Grace Period	Período de carência
Late (16 - 30 days)	Atraso entre 16 e 30 dias
Default	Inadimplente

Fonte: Elaboração própria.

Uma vez que queremos modelar o comportamento de empréstimos inadimplentes ou não, é necessário remover da base de dados empréstimos que ou não alcançaram a capacidade

Figura 9 – Descrição - Loan Status



Fonte: Elaboração própria.

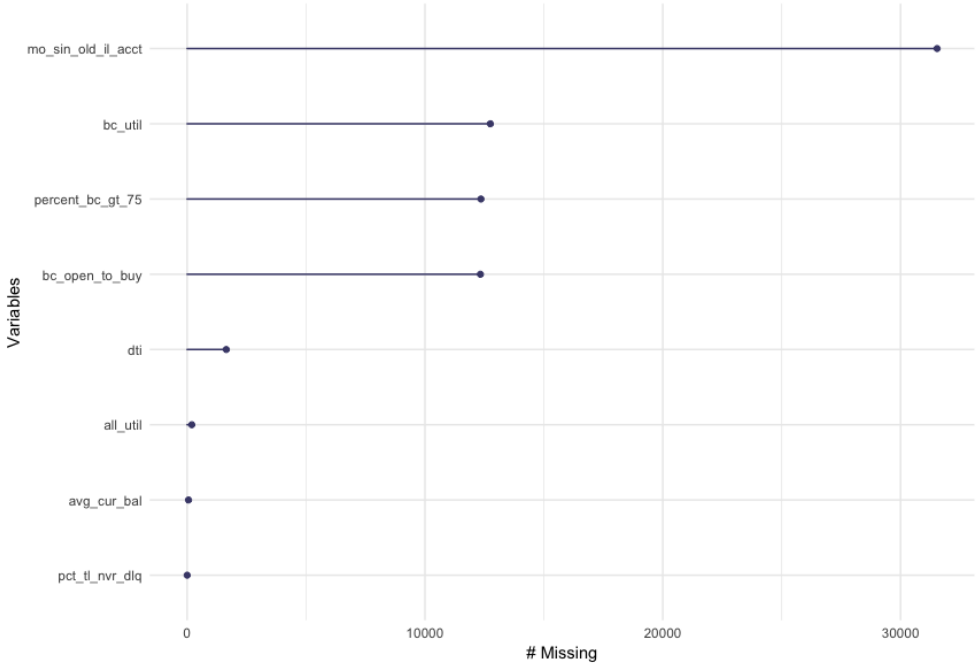
de se tornar inadimplentes (apresentar atrasos acima de 120 dias) ou ainda estão no período de carência, onde nenhum pagamento sequer pôde ser realizado. Filtramos a base para que a variável *loan_status* contenha apenas as observações das seguintes categorias: *Charged Off, Default, Fully Paid, Current*. Realizado esse filtro, podemos construir a nossa variável dependente:

$$Y_i = \begin{cases} 1, \text{se loan\_status = Charged Off ou loan\_status = Default} \\ 0, \text{se loan\_status = Fully Paid} \end{cases}$$

3.3 Tratamentos

O primeiro tratamento a ser realizado é a remoção de variáveis que possuem muitos campos preenchidos com *NA - Not Available* e para isso podemos fazer um gráfico que exibe quantas observações de cada variável são *NA*, expomos esse gráfico na Figura 10. Das 8 variáveis que apresentam observações faltantes, decidimos manter apenas dti, all_util, avg_cur_bal e pct_tl_nvr_dlq.

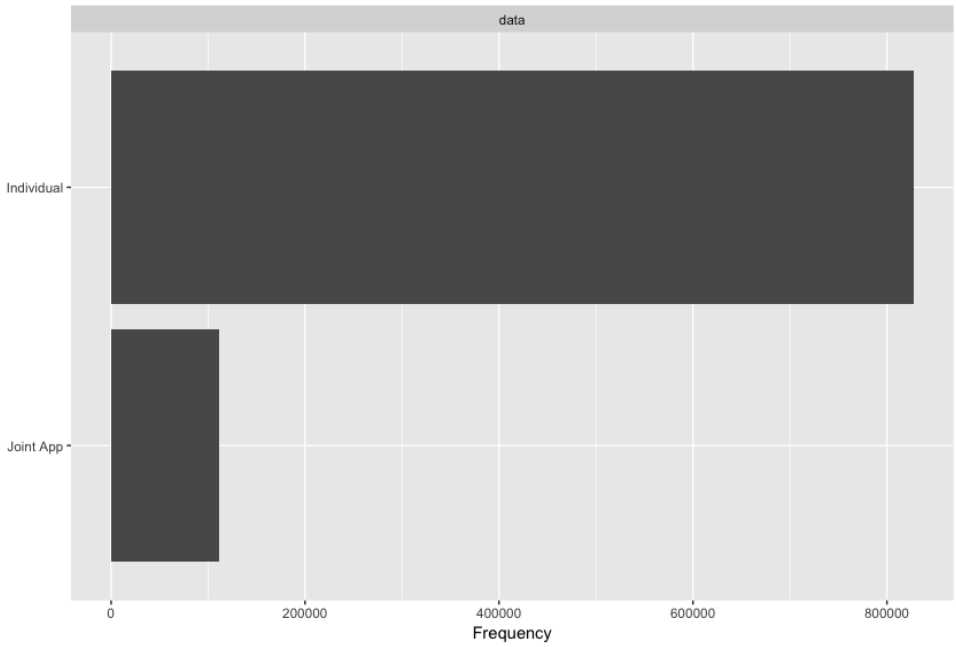
Figura 10 – Missing Map



Fonte: Elaboração própria.

Na Lending Club, as solicitações de crédito podem ser realizadas de forma individual ou conjuntamente com o conjunge, porém, essas aplicações conjuntas representam menos de 12% de todas as concessões realizadas no período aqui analisado conforme exposto na Figura 11. Uma vez que as aplicações conjuntas possuem particularidades operacionais que mitigam o risco da operação, vamos excluir essas observações do estudo, aqui realizando o uso dos pacote `dplyr`, desenvolvido por Wickham et al. 2019 e `naniar`, desenvolvido por Tierney et al. 2019.

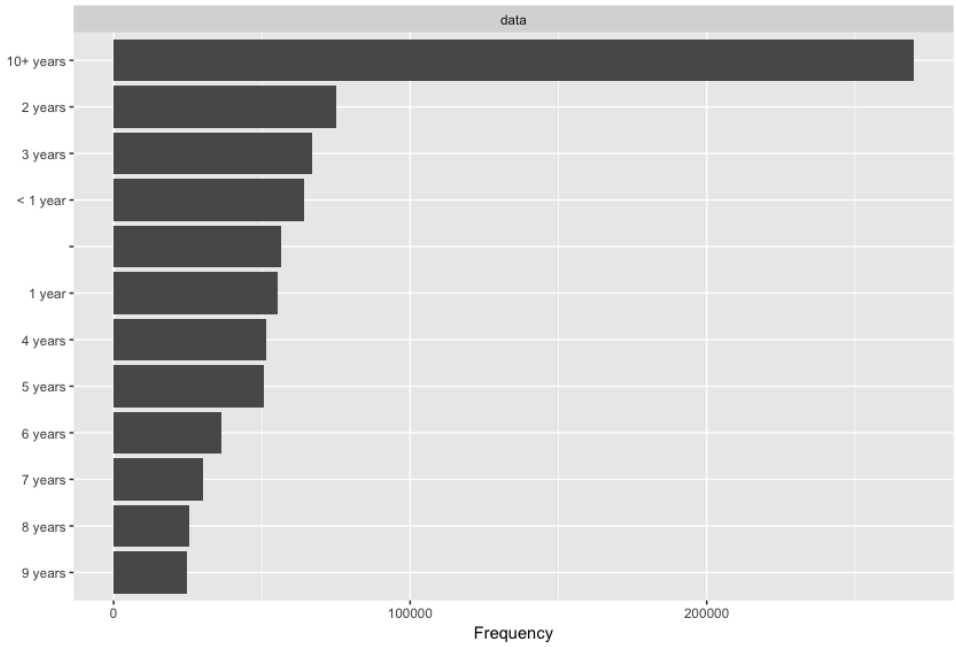
Figura 11 – Distribuição - Tipo de Aplicação



Fonte: Elaboração própria.

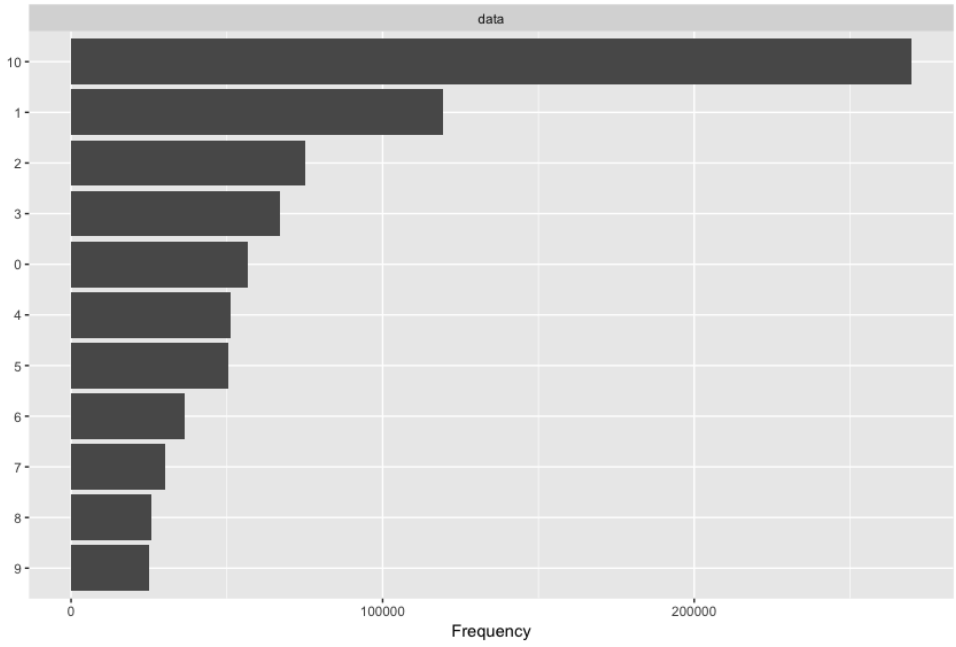
A variável que determina há quanto tempo o tomador está empregado na empresa em que trabalha atualmente (*emp_length*) está armazenada como texto em categorias e precisamos convertê-la para uma variável discreta como mostramos nas Figuras 12 e 13, aqui um valor vazio significa que o tomador está desempregado, portanto, possui 0 anos de experiência no emprego atual.

Figura 12 – Distribuição - Tempo de Emprego



Fonte: Elaboração própria.

Figura 13 – Discretização - Tempo de Emprego



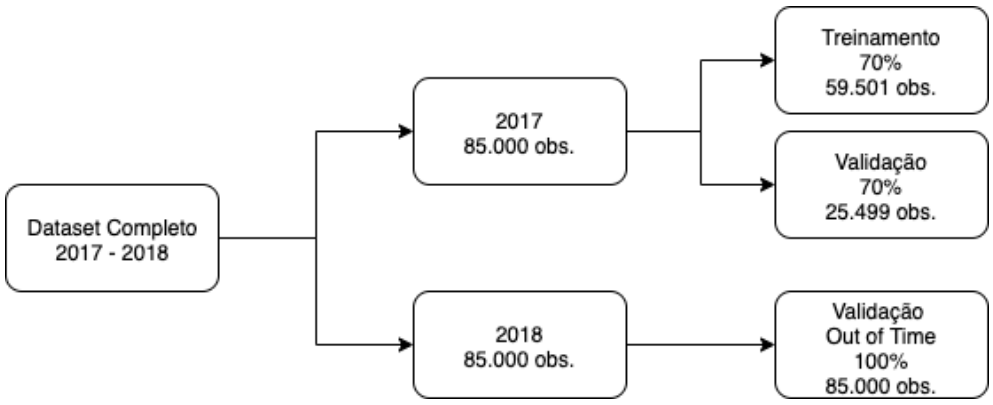
Fonte: Elaboração própria.

3.4 Separação - Treino, Validação e Validação Out of Time

Definimos aqui a separação entre as bases de treino (para estimação do vetor β), validação (performance do modelo fora da amostra de treino e *Cross Validation*) e teste (validação fora do

tempo) conforme exposto no diagrama da figura 14.

Figura 14 – Diagrama - Separação entre Treino, Validação e Teste



Fonte: Elaboração própria.

Considerando todos os empréstimos concedidos ao longo de 2017, 70% foi alocado para o treino do modelo, ou seja, a estimação dos coeficientes β e 30% para a validação (contemplando também a validação cruzada). Os empréstimos de 2018 foram todos alocados para a validação *out of time*, ou seja, emprréstimos que não participam do mesmo período que os que foram usados para o treino e validação dos parâmetros obtidos durante a estimação.

3.5 Estimação dos coeficientes

Uma vez realizado o tratamento da base de dados e a criação da variável dependente, podemos estimar o modelo de regressão logística a partir das 59501 observações disponíveis na base de treinamento, conforme descrito na Figura 14, para realizar a estimação, utilizaremos as funções do pacote `caret` desenvolvido por Wing et al. 2019. As variáveis que participaram do modelo estão dispostas na Tabela 5 e seus respectivos coeficientes estimados estão dispostos na Tabela 6 e vale ressaltar que o grupo de referência corresponde a variável omitida grade $=$ A. Os códigos fundamentais utilizados na configuração do ambiente de estimação e na geração das métricas de validação do modelo estimado estão disponíveis no Apêndice B.

Tabela 5 – Descrição das Variáveis

Variável	Tipo	Valores
Default	Dependente	{0, 1}
dti	Contínua	
installment	Contínua	
fico_range_high	Contínua	
fico_range_low	Contínua	
mort_acc	Contínua	
annual_inc	Contínua	
inq_last_6mths	Discreta	[0;12]
num_rev_accts	Discreta	[0;9]
acc_open_past_24mths	Discreta	[0;10]
grade	Discreta	{A, B, C, D, E, F}
emp_length	Discreta	[0;10]
term	Discreta	{36 months; 60 months}

Fonte: Elaboração própria.

Tabela 6 – Coeficientes do modelo estimado

	Regressão Logística
(Intercept)	5.130921379
	(3.144699638)
dti	-0.008560924^{***}
	(0.001177505)
installment	-0.000703709^{***}
	(0.000037560)
fico_range_high	-1.674042652^{*}
	(0.786937705)
fico_range_low	1.678954141^{*}
	(0.786957963)
mort_acc	0.057337809^{***}
	(0.005932815)
‘term60 months‘	−0.002169589
	(0.021851985)
annual_inc	0.000002559^{***}
	(0.000000226)
inq_last_6mths	-0.152713837^{***}
	(0.010920250)
num_rev_accts	0.003531696^{**}
	(0.001323632)
acc_open_past_24mths	-0.050376991^{***}
	(0.003143933)
gradeB	-0.677278383^{***}
	(0.038965541)
gradeC	-1.229864621^{***}
	(0.040076104)
gradeD	-1.549224331^{***}
	(0.044670170)
gradeE	-1.875138935^{***}
	(0.053588453)
gradeF	-2.194692401^{***}
	(0.076307005)
gradeG	-2.263325350^{***}
	(0.095057309)
emp_length	0.025955559^{***}
	(0.002432190)
AIC	71760.766657667
BIC	71922.654128833
Log Likelihood	-35862.383328833
Deviance	71724.766657667
Num. obs.	59501

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

4 VALIDAÇÃO DO MODELO

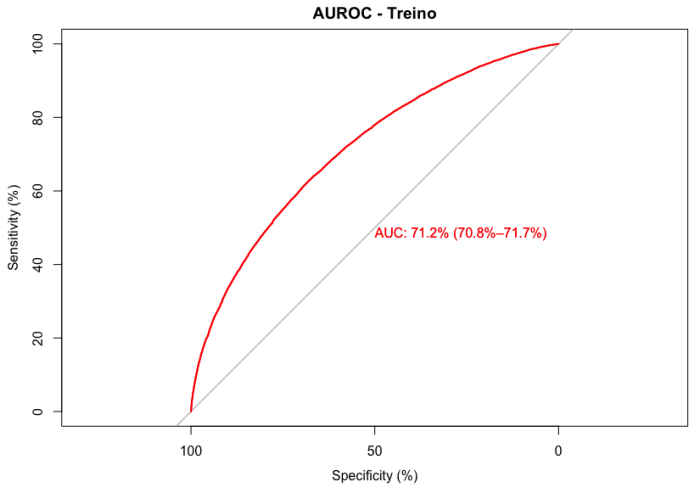
Uma vez estimado o modelo, podemos verificar qual a performance do mesmo utilizando as métricas convencionais (AIC, BIC, Log Likelihood) expostas na Tabela 6, porém, em problemas de classificação de risco de crédito as métricas relevantes derivam da necessidade de diferenciar bons tomadores de maus tomadores e por isso vamos expor as métricas AUROC e KS em cada um dos segmentos de estimação analisados.

4.1 AUROC

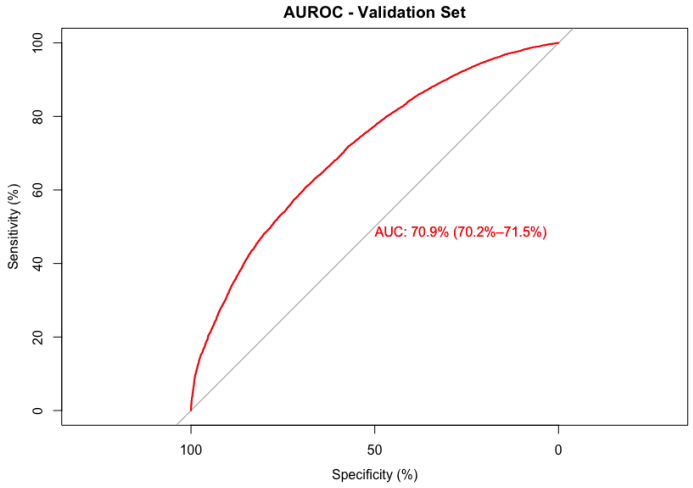
O modelo apresentou uma performance bastante satisfatória na base de dados de treinamento de acordo com a Tabela 2 uma vez que gerou uma AUROC igual a 71,2% (entre 70,8% e 71,7% considerando um intervalo de confiança de 95%). Para facilitar a visualização da performance, exibimos todas as métricas de performance no Apêndice C.

Na base de dados de validação o modelo manteve seu desempenho satisfatório, alcançando uma AUROC igual a 70,9%, porém, precisamos garantir que esse resultado é uniforme ao longo de toda a base de validação e para isso realizamos a *Cross Validation* utilizando 5 *folds*(dobras). Os resultados dessa técnica de validação, expostos na Figura 16 demonstram que em todas as dobras a performance do modelo foi satisfatória uma vez que toda as AUROC estão acima de 70%.

Figura 15 – AUROC - Treino e Validação



(a) Treino



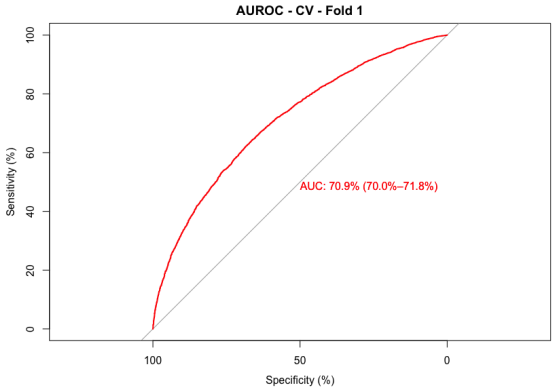
(b) Validação

Fonte: Elaboração própria.

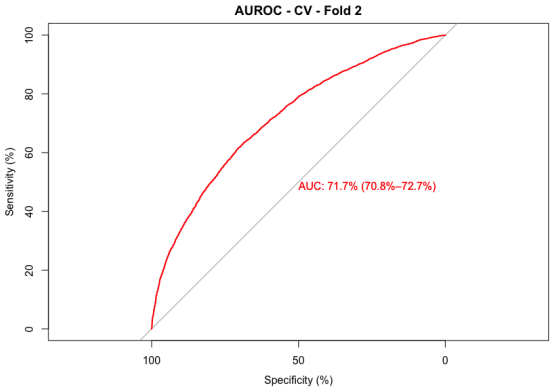
Para garantir a estabilidade do modelo, precisamos avaliar a sua performance em dados provenientes de um período temporal diferente dos que foram usados tanto no treino quanto na validação. Uma vez que utilizamos os dados de 2017 para obter os parâmetros e a performance fora da amostra, podemos utilizar as observações dos empréstimos concedidos ao longo de 2018 para realizar a validação *Out of Time*.

A Figura 17 demonstra que a AUROC do modelo em dados de 2018 se localiza no intervalo entre 70,4% e 71,6%, ou seja, o modelo performa de forma satisfatória mesmo ao ser apresentado de dados fora do tempo que foi treinado e sua performance permanece estável em empréstimos de 1 ano a frente em relação a 2017.

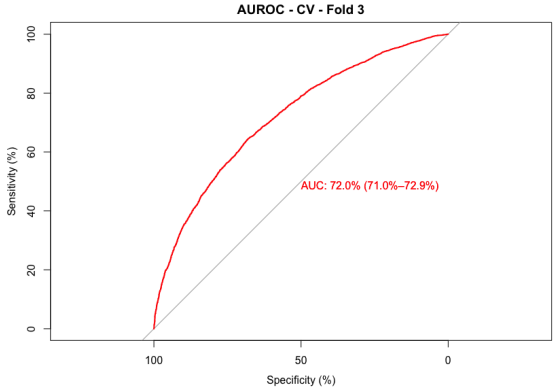
Figura 16 – AUROC - Cross Validation



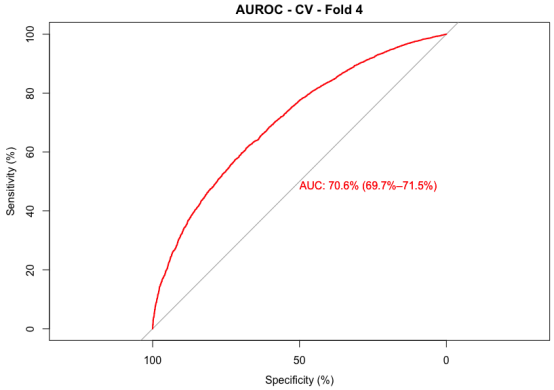
(a) Fold 1



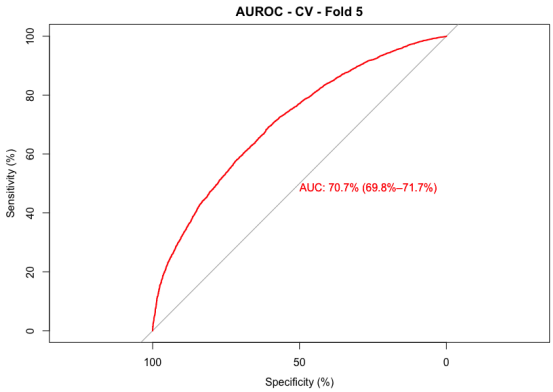
(b) Fold 2



(c) Fold 3



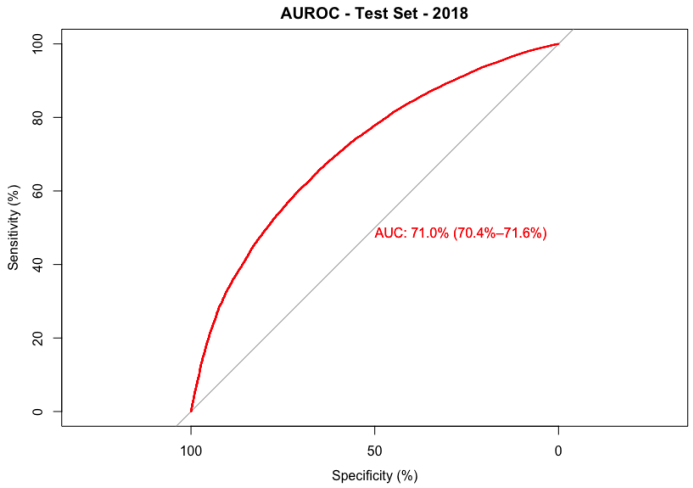
(d) Fold 4



(e) Fold 5

Fonte: Elaboração própria.

Figura 17 – AUROC - Out of Time



Fonte: Elaboração própria.

4.2 KS

Analisando a estatística KS de cada um dos segmentos avaliados, podemos afirmar que todas as estatísticas KS estão em um patamar de boa discriminação de acordo com a Tabela 3 com exceção das dobras 4 e 5 do *Cross Validation* e a amostra de validação que resultou em um KS aceitável e igual a 0.2934. Vale ressaltar que aqui não se trata de um p-valor e sim da estatística KS em si, conforme descrito na equação (2.14).

Tabela 7 – Kolmogorov-Smirnov

Segmento	KS
Treino	0.3065
CV - Fold 1	0.3056
CV - Fold 2	0.3192
CV - Fold 3	0.3217
CV - Fold 4	0.2933
CV - Fold 5	0.2925
Validação	0.2934
Teste - 2018	0.3075

Fonte: Elaboração própria.

5 CONSIDERAÇÕES FINAIS

Nesta monografia apresentamos de forma introdutória os conceitos básicos de regressão logística e sua implementação computacional além dos conceitos necessários para a avaliação de performance e validação (ROC, AUROC e KS) como ferramentas alternativas na escolha de modelos. Aplicamos as técnicas analisadas no segmento de classificação de risco de crédito utilizando o software R em todas as etapas práticas e obtivemos um modelo de predição satisfatório de acordo com critérios objetivos e comparáveis entre si, podemos afirmar então que os resultados desse trabalho comprovam a capacidade da regressão logística de prover estimativas suficientes para classificação de risco de crédito e seu posterior uso para precificação e gerenciamento de carteiras de crédito de varejo.

6 REFERÊNCIAS

ARNOLD, T. B.; EMERSON, J. W. Nonparametric goodness-of-fit tests for discrete null distributions. *R Journal*, v. 3, n. 2, 2011. Citado na página 24.

COMMITTEE, B. et al. *Update on work of the accord implementation group related to validation under the Basel II framework*. [S.l.], 2005. Citado na página 12.

DOBSON, A. J.; BARNETT, A. *An introduction to generalized linear models*. [S.l.]: Chapman and Hall/CRC, 2008. Citado na página 17.

DUTANG, C. Some explanations about the iwls algorithm to fit generalized linear models. 2017. Citado na página 17.

FAWCETT, T. An introduction to roc analysis. *Pattern recognition letters*, Elsevier, v. 27, n. 8, p. 861–874, 2006. Citado 5 vezes nas páginas 19, 20, 21, 22 e 24.

GOURIEROUX, C. *Econometrics of qualitative dependent variables*. [S.l.]: Cambridge university press, 2000. Citado 4 vezes nas páginas 13, 15, 16 e 17.

GREENE, W. H. A statistical model for credit scoring. NYU Working Paper No. EC-92-29, 1992. Citado 2 vezes nas páginas 11 e 12.

KOHAVI, R. et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: MONTREAL, CANADA. *Ijcai*. [S.l.], 1995. v. 14, n. 2, p. 1137–1145. Citado na página 26.

LENDING Club Statistics. 2019. Disponível em: <https://www.lendingclub.com/info/demand-and-credit-profile.action>. Acesso em: 23 abr. 2019. Citado na página 27.

NELDER, J. A.; WEDDERBURN, R. W. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, Wiley Online Library, v. 135, n. 3, p. 370–384, 1972. Citado na página 13.

OLIVEIRA, J. G.; ANDRADE, F. W. Comparação entre medidas de performance de modelos de credit scoring. *Tecnologia de Crédito*, v. 33, p. 35–47, 2002. Citado 2 vezes nas páginas 24 e 25.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2018. Disponível em: <https://www.R-project.org/>. Citado na página 27.

TIERNEY, N. et al. *naniar: Data Structures, Summaries, and Visualisations for Missing Data*. [S.l.], 2019. R package version 0.4.2. Disponível em: <https://CRAN.R-project.org/package=naniar>. Citado na página 29.

WICKHAM, H. et al. *dplyr: A Grammar of Data Manipulation*. [S.l.], 2019. R package version 0.8.1. Disponível em: <https://CRAN.R-project.org/package=dplyr>. Citado na página 29.

WING, M. K. C. from J. et al. *caret: Classification and Regression Training*. [S.l.], 2019. R package version 6.0-84. Disponível em: <https://CRAN.R-project.org/package=caret>. Citado na página 32.

A DICIONÁRIO DOS DADOS

Tabela 8 – Dicionário dos Dados

Nome da Variável	Descrição
acc_open_past_24mths	Número de contas abertas nos últimos 24 meses.
addr_state	O estado de residência (UF)
all_util	Limite de crédito em todas as contas sob titularidade do tomador.
annual_inc	A renda anual auto declarada pelo tomador durante o registro.
avg_cur_bal	Média do saldo atual de todas as contas sob titularidade do tomador.
bc_open_to_buy	Saldo não utilizado em cartões de crédito.
bc_util	Proporção entre o saldo devedor no cartão de crédito e os demais limites.
chargeoff_within_12_mths	Número de eventos de cobrança nos últimos 12 meses.
collections_12_mths_ex_med	Número de eventos de cobrança nos últimos 12 meses (exceto médicas)
delinq_2yrs	Número de eventos de atraso superior a 30 dias nos últimos 2 anos.
delinq_amnt	Soma dos saldos cuja conta está em estado de delinquência
dti	Proporção entre o pagamento mensal de dívida e a renda mensal
earliest_cr_line	Mês de abertura da linha de crédito mais antiga do tomador.
emp_length	Número de anos que o tomador está empregado no emprego atual.
fico_range_high	Limite superior do intervalo do score FICO.
fico_range_low	Limite inferior do intervalo do score FICO.
funded_amnt	Valor total financiado no empréstimo.
grade	Score de crédito calculado pela LendingClub
home_ownership	Tipo de residência (alugado, próprio etc) declarado no registro
inq_last_12m	Número de solicitações de crédito nos últimos 12 meses
inq_last_6mths	Número de solicitações de crédito nos últimos 6 meses
installment	Valor mensal da parcela do empréstimo
loan_amnt	Valor do empréstimo solicitado pelo tomador.
loan_status	Status do empréstimo
max_bal_bc	Saldo máximo devido em todas as contas de crédito rotativo.
mo_sin_old_il_acct	Meses desde a última abertura de conta para pagamentos de parcelas
mo_sin_old_rev_tl_op	Meses desde a última abertura de contas de crédito rotativo
mo_sin_rcnt_rev_tl_op	Meses desde a mais recente abertura de conta de crédito rotativo
mo_sin_rcnt_tl	Meses desde a mais recente abertura de conta
mort_acc	Número de contas de hipoteca sob titularidade do tomador
num_accts_ever_120_pd	Número de contas que apresentaram atraso acima de 120 dias
num_actv_bc_tl	Número de contas bancárias atualmente ativas
num_rev_accts	Número de contas de crédito rotativo
open_acc	Número de linhas de crédito abertas no histórico do tomador.
open_acc_6m	Número de contas abertas nos últimos 6 meses.
open_il_12m	Número de financiamentos contratados nos últimos 12 meses.
open_il_24m	Número de financiamentos contratados nos últimos 24 meses.
open_act_il	Número de financiamentos ativos.
open_rv_12m	Número de contas rotativas abertas nos últimos 12 meses
open_rv_24m	Número de contas rotativas abertas nos últimos 24 meses
pct_tl_nvr_dlq	Número de contas rotativas que nunca apresentaram atrasos
pub_rec	Número de registros derrogatórios públicos
pub_rec_bankruptcies	Número de falências públicas declaradas pelo tomador
purpose	Motivo de solicitação do empréstimo
term	Número de parcelas do empréstimo (36 ou 60 meses)
application_type	Tipo de aplicação (conjunta ou individual)

Fonte: Elaboração própria.

B CÓDIGO NO R

```
ctrl <- trainControl(
  method = "cv", # Cross Validation
  number = 5, # 5 folds
  savePredictions = TRUE, # Salva o resultado de cada modelo
  classProbs = TRUE,
  summaryFunction = twoClassSummary,
  verboseIter = TRUE
)


LogitModel <- train(default ~ dti +
installment +
fico_range_high +
fico_range_low +
mort_acc +
term +
annual_inc +
inq_last_6mths +
num_rev_accts +
acc_open_past_24mths +
grade +
dti +
emp_length,
data = LC_2017_Train,
family = "binomial",
method = "glm",
metric = "ROC",
trControl = ctrl)


# Validation - ROC

plot.roc(LC_2017_Train$default,
     predict.train(LogitModel,
     newdata = LC_2017_Train,
     type = "prob")$Bad,
```

```
    col = "red",
    main = "AUROC_Treino",
    percent = TRUE,
    print.auc = TRUE,
    ci = TRUE)

# Validation - KS

ks_stat(LC_2017_Train$default,
predict.train(LogitModel, newdata = LC_2017_Train,
type = "prob")$Bad)
```

C MÉTRICAS DE PERFORMANCE

Tabela 9 – Métricas de Performance - AUROC

Segmento	AUROC	Limite Inferior	Limite Superior
Treino - 2017	71.2%	70.8%	71.7%
Validação - 2017	70.9%	70.2%	71.5%
Cross Validation - Fold 1	70.9%	70.0%	71.8%
Cross Validation - Fold 2	71.7%	70.8%	72.7%
Cross Validation - Fold 3	72.0%	71%	72.9%
Cross Validation - Fold 4	70.6%	69.7%	71.5%
Cross Validation - Fold 5	70.7%	69.8%	71.7%
Out of Time - 2018	71.0%	70.4%	71.6%

Fonte: Elaboração própria.