Instituto de MATEMÁTICA E ESTATÍSTICA UFRGS

DEPARTAMENTO DE ESTATÍSTICA
E-mail: dest@mat.ufrgs.br

Trabalho de Conclusão de Curso

Regressão Logística Geograficamente Ponderada na Análise de Risco de Crédito

Raquel Rossi Ferreira

27 de maio de 2021

Raquel Rossi Ferreira

Regressão Logística Geograficamente Ponderada na Análise de Risco de Crédito

Trabalho de Conclusão apresentado à comissão de Graduação do Departamento de Estatística da Universidade Federal do Rio Grande do Sul, como parte dos requisitos para obtenção do título de Bacharel em Estatística.

Orientadora: Profa. Dr[a]. Lisiane Priscila Roldão Selau

Co-orientadora: Profa. Dr[a]. Márcia Helena Barbian

Porto Alegre
Maio de 2021

Raquel Rossi Ferreira

Regressão Logística Geograficamente Ponderada na Análise de Risco de Crédito

Este Trabalho foi julgado adequado para obtenção dos créditos da disciplina Trabalho de Conclusão de Curso em Estatística e aprovado em sua forma final pela Orientadora e pela Banca Examinadora.

Orientadora:_____
Profa. Dr[a]. Lisiane Priscila Roldão Selau, UFRGS
Doutora pela Universidade Federal do Rio Grande do Sul, Porto Alegre, RS

Banca Examinadora:

Profa. Dr[a]. Márcia Helena Barbian, UFMG,
Doutora pela Universidade Federal de Minas Gerais, Belo Horizonte, MG

Mariana Nolde Pacheco Detoni,
Bacharel em Estatística pela UFRGS - SICREDI

Porto Alegre
Maio de 2021

"Time flies. Time waits for no man. Time heals all wounds. All any of us wants is more time. Time to stand up. Time to grow up. Time to let go. Time".
Meredith Grey, Grey's Anatomy

Agradecimentos

Gostaria de agradecer a minha família por todo apoio durante o período de graduação, em especial a minha mãe que passou junto comigo por todo o processo do vestibular até ver o meu nome no listão e que durante o curso comemorou cada aprovação nas disciplinas.

Ao Ronaldo, além de ser um ótimo irmão, nunca se importou em emprestar o computador (porque o dele era mais rápido) para que eu pudesse compilar meus códigos no R e estudar para as provas. Além de sempre deixar o café pronto todas as manhãs.

Aos meus colegas de curso: Franciele, Gabriela, Juliana, Maicon e Maitê. Franciele e Maitê obrigada por todo o apoio durante a graduação explicando aquelas matérias que pareciam impossíveis, não teria conseguido chegar até aqui sem a ajuda de vocês. A Juliana por ter sido a melhor monitora da história da universidade e uma grande amiga. Ao Maicon e a Gabriela por serem os melhores parceiros de bar no sábado a noite.

A Ana, Bruna e o Wylliam por fazerem o meu estágio muito mais divertido, obrigada por escutarem as minhas reclamações do dia a dia, por me fazerem rir e me apoiarem nas minhas decisões (por mais malucas que elas fossem). Vocês são os melhores amigos que alguém poderia ter.

As minhas orientadoras professoras Lisiane Selau e Márcia Barbian, muito obrigada por todo o suporte, motivação e ensinamentos.

A todos os professores do Instituto de Matemática e Estatística da UFRGS por todo o aprendizado oferecido durante a graduação.

Por fim, um grande agradecimento a UFRGS, pela oportunidade de ensino, aprendizado e crescimento profissional e pessoal.

Resumo

O mercado brasileiro de crédito se popularizou nos últimos anos e com isso empresas buscam formas de aumentar a capacidade de previsão de seus modelos na hora de conceder crédito. Há diversas técnicas que são amplamente utilizadas pelo setor financeiro para a construção dos modelos de previsão, a mais usada é a Regressão Logística. Recentemente estudos mostraram que o uso da Regressão Logística Geograficamente Ponderada tem um desempenho melhor para prever o comportamento padrão de risco quando comparada com métodos de regressão não espacial. Sendo assim, este trabalho introduz o modelo de Regressão Logística Geograficamente Ponderada e compara com os resultados do modelo tradicional usado pelo mercado financeiro, a Regressão Logística. Com os objetivos de verificar se existe influência do espaço geográfico no risco de crédito de clientes de uma rede de farmácias e verificar se a técnica de Regressão Logística Geograficamente Ponderada tem maior capacidade de previsão que a Regressão Logística na avaliação de risco de crédito. Os modelos foram desenvolvidos com uma base de dados de treino de 2014 clientes, e foram avaliados em um conjunto de teste de 515 clientes. Os modelos foram analisados com base em quatro indicadores: percentual de acerto, área abaixo da curva ROC, critério de informação de Akaike e teste KS. Neste estudo a técnica de Regressão Logística Geograficamente Ponderada obteve resultados superiores a Regressão Logística no desenvolvimento de modelos de *Credit Scoring*. No entanto, os resultados de ambos os modelos foram bem próximos em termos de capacidade de previsão, mas o modelo de Regressão Logística Geograficamente Ponderada teve resultados pouco melhores em relação a Regressão Logística.

Palavras-Chave: Credit Scoring, Regressão Logística Geograficamente Ponderada, Regressão Logística.

Abstract

The Brazilian credit market has become more popular in recent years and as a result, companies are looking for ways to increase the predictive capacity of their models when lending credit. There are several techniques that are widely used by the financial sector to build predictive models, the most used is Logistic Regression. Recently studies have shown that the use of Geographically Weighted Logistic Regression performs better to predict the standard risk behavior when compared to non-spatial regression methods. Therefore, this work introduces the Geographically Weighted Logistic Regression model and compares it with the results of the traditional model used by the financial market, the Logistic Regression. The objectives are to verify whether there is an influence of the geographic space on the credit risk of customers of a pharmacy chain and to verify whether the Geographical Weighted Logistic Regression technique has greater predictive capacity than Logistic Regression in the assessment of credit risk. The models were developed based on a training set of 2014 clients, and they were evaluated in a validation set of 515 clients. The models were analyzed based on four indicators: hit percentage, area below the ROC curve, Akaike information criterion and KS test. In this study, the Geographical Weighted Logistic Regression technique obtained superior results when compared to Logistic Regression in the development of *Credit Scoring* models. However, the results of both models were very close in terms of predictive capacity, but the Geographically Weighted Logistic Regression model had slightly better results compared to Logistic Regression.

Keywords: Credit Scoring, Geographically Weighted Logistic Regression, Logistic Regression.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

Na área de risco de crédito, o uso de técnicas para reconhecer quais consumidores serão bons ou maus pagadores é tarefa importante e muitas vezes difícil, pois se um cliente for classificado de maneira incorreta pode causar prejuízos a instituição que concede crédito (classificar um cliente mau como bom) ou então impedir a instituição de obter ganhos (classificar um cliente bom como mau). Os avanços tecnológicos associados ao desenvolvimento de métodos quantitativos colaboraram para a criação de muitas técnicas para a mensuração de risco de crédito (Silva, 2011), os denominados modelos de *Credit Scoring*.

Para a mensuração do risco de crédito, existem várias técnicas utilizadas pelas instituições financeiras, como por exemplo, Sistemas especialistas, Sistemas de ratings, Sistemas de escores para crédito e também existem novas abordagens para os modelos de mensuração, como a metodologia VaR *(Value at Risk)*, e a metodologia *Credit Metrics* (Silva, 2011).

Os modelos de análise para concessão de crédito conhecidos como modelos de *Credit Scoring* são baseados em dados históricos da base de clientes existentes, com base nessas informações o modelo avalia se um futuro cliente será um bom ou mau pagador. Por envolverem menor custo, agilidade, objetividade e poder preditivo na decisão da concessão de crédito, os modelos de *Credit Scoring* se popularizaram e são amplamente utilizados pelo setor financeiro (Hand & Henley, 1997; Albuquerque & Silva, 2017).

Modelos que avaliam o crédito são de extrema importância para uma instituição financeira, porém nenhum modelo consegue uma precisão absoluta. Por isso há o interesse de se analisarem diferentes tipos de modelos e apontar quais apresentam maior precisão para cada caso.

Muitas pesquisas sobre diferentes métodos de classificação para modelos de *Credit Scoring* já foram realizadas na literatura, sendo a Regressão Logística a considerada padrão no setor financeiro, como indicado na pesquisa de Lessmann *et al.* (2015). A Regressão Logística é uma técnica estatística que busca explicar a relação entre uma variável dependente binária e um conjunto de variáveis preditoras (Hosmer & Lemeshow, 2000).

Neste estudo, propõe-se o uso de geoprocessamento como uma abordagem para melhorar a acurácia dos modelos de *Credit Scoring*, buscando verificar se fatores que influenciam o risco de crédito diferem de acordo com a localização geográfica do consumidor.

A localização geográfica e sua relação com o risco de crédito é tema de alguns estudos publicados. Dentre os mais recentes, Stine (2011) analisa a evolução da

inadimplência do crédito imobiliário em condados dos Estados Unidos entre 1993 e 2010, tendo encontrado evidências da existência de correlação espacial entre as taxas de inadimplência daqueles condados (Albuquerque & Silva, 2017).

Fernandes & Artes (2016) aplicaram a metodologia *Ordinary Kriging* para criar uma variável que reflete o risco espacial e aplicaram a técnica de Regressão Logística para verificar a existência de correlação espacial na inadimplência de pequenas e médias empresas. Os autores desenvolveram modelos com e sem a variável de risco espacial e confirmaram que a inclusão dessa variável melhora o desempenho dos modelos de *Credit Scoring* (Albuquerque & Silva, 2017).

Uma vantagem da aplicação da técnica de Regressão Logística Geograficamente Ponderada (RLGP) em relação a outras técnicas para este fim é a estimação de um modelo para cada região do estudo, possibilitando que as estimativas desses modelos sejam distintas em relação ao risco de crédito, visto que regiões de estudo diferentes podem possuir riscos diferentes (Atkinson *et al.*, 2003). Destaca-se ainda, que a RLGP é mais adequada principalmente em situações na qual características locais podem diferenciar melhor o risco de crédito dos clientes que residem naquela região e, consequentemente, gerar ganhos financeiros para a instituição. Esse estudo tem duas principais hipóteses:

- Verificar se existe influência do espaço geográfico no risco de crédito de clientes de uma rede de farmácias;
- Verificar se a técnica de Regressão Logística Geograficamente Ponderada tem maior capacidade de previsão que a Regressão Logística na avaliação de risco de crédito.

A principal contribuição teórica deste estudo é ilustrar a aplicação da técnica de Regressão Logística Geograficamente Ponderada na construção de modelos de *Credit Scoring* por meio do *Software* R. Não foram encontrados estudos que utilizaram a técnica RLGP com a utilização de algortimos do *Software* R. Já a contribuição prática do tema é que a técnica de Regressão Logística Geograficamente Ponderada é de extrema importância como meio de análise de concessão de crédito, pois considera as características regionais específicas que podem alterar o risco e, por consequência, auxilia as instituições a melhor administrar seus ganhos e prejuízos.

A capacidade de previsão dos modelos será avaliada pelas seguintes métricas: percentual de acerto, área abaixo da curva ROC, Critério de Informação de Akaike (AIC) e medida do teste Kolmogorov-Smirnov (KS).

Esse estudo irá utilizar uma fonte de dados secundários com informações sobre os clientes de uma rede de farmácias com crediário próprio e unidades em todo o Rio Grande do Sul, obtidos no estudo de (Selau, 2008). Estes dados referem-se a clientes incluidos no período de dezembro de 2005 a junho de 2006.

Nesse sentido, o estudo tem como principal resultado o entendimento se o fator geográfico influencia no risco de crédito, utilizando modelagem estatística.

2 Metodologia

O estudo é desenvolvido através das etapas descritas a seguir:

1. Revisão bibliográfica
2. Organização e limpeza dos dados
3. Identificação de uma função para RLGP
4. Modelagem da RL e RLGP
5. Comparação dos resultados entre RL e RLGP
6. Discussão dos resultados

Primeiramente, será realizada uma breve revisão teórica sobre a importância da RLGP na análise de risco de crédito, além de entender mais sobre as medidas estatísticas de percentual de acerto, área abaixo da curva ROC, Critério de Informação de Akaike (AIC) e teste KS. Na Fase 2 será feita a definição das variáveis que irão compor o estudo, validar consistência e preenchimento dos dados e separar parte da amostra para teste dos modelos. Na Fase 3 será utilizada uma função para o uso da RLGP. Para modelagem de dados, Fase 4, o estudo seguirá as etapas propostas por (Sicsú, 2010).

O desenvolvimento de um modelo de *Credit Scoring* compreende as seguintes etapas:

1. Planejamento e definições
2. Identificação de variáveis potenciais
3. Planejamento amostral
4. Aplicação de metodologia estatística para determinação do score
5. Validação e verificação de performance do modelo estatístico

Na Fase 5, será verificado o desempenho dos modelos desenvolvidos através da comparação dos resultados das métricas utilizadas (percentual de acerto, área abaixo da curva ROC, Critério de Informação de Akaike e teste KS).

Por fim, na Fase 6, os resultados serão compilados e uma análise crítica dos modelos será realizada para encerramento do estudo.

Como citado anteriormente a base de dados utilizada será de uma rede de farmácias que oferece um cartão de crédito para os clientes de forma que seja mais fácil realizar o pagamento das compras.

2.1 Banco de dados

Para a empresa em estudo, atraso de até 30 dias define o cliente como bom, atrasos superiores a 60 dias define o cliente como mau. Atrasos na faixa de 31 a 60 dias são clientes classificados como indefinidos e por isso são excluídos do processo de modelagem, para que seja possível conseguir um maior poder de discriminação entre clientes bons e maus (Selau, 2008). Existe também um quarto grupo de clientes composto por aqueles que ainda não utilizaram o cartão no período do estudo, sendo assim eles foram classificados como clientes sem uso.

Utilizou-se apenas os clientes que foram classificados como bons e maus, em um primeiro momento a amostra do total de clientes é de 11681.

As informações para a criação do modelo foram disponibilizadas pela empresa, as variáveis preditoras são provenientes da proposta preenchida pelos clientes no momento da solicitação do crédito. Inicialmente foram consideradas as 16 variáveis listadas na Tabela (2.1).

Realizou-se uma avaliação geral do preenchimento dos dados para eliminar a inconsistência, *outliers* ou *missing*. Nas variáveis Idade e Tempo de serviço, foi constatado a ocorrência de valores negativos e também casos de clientes com menos de 18 anos, o que demonstra erros de preenchimento dos campos no momento do cadastro. Na variável CEP residencial, observou-se a presença de valores gerais zerados ou que pertenciam a outros estados. Foi decidido que esses casos deveriam ser excluídos fazendo com que a amostra resultante fosse de 11389 clientes.

As amostras de treino e teste foram separadas de formas aleatória na proporção de 80% e 20% respectivamente. A ideia desta separação é utilizar uma parcela dos dados para a criação dos modelos, e a outra, para verificar o desempenho do modelo em uma amostra diferente. Sendo assim, a amostra de treino ficou composta de 5509 clientes bons e 3601 maus, com um total de 9110 clientes. Já a amostra de teste foi formada de 1378 bons e 901 maus, totalizando 2279 clientes.

A decisão de se categorizar as variáveis preditoras foi a grande quantidade de variáveis a serem modeladas e com intensa variabilidade, o que poderia causar a rejeição de variáveis que poderiam ser importantes e seriam excluídas pela modelagem, pois seus coeficientes poderiam ser considerados não significativos pelo modelo.

Inicialmente, foi feita uma análise da escolha das variáveis através do cálculo do Risco Relativo (RR), onde o percentual de bons clientes foi dividido pelo percentual de maus para cada um dos atributos. Por meio deste cálculo foram excluídas duas variáveis do processo de análise (Tipo Renda e Crédito 3°s), pois ambas apresentaram um Risco Relativo próximo de 1. A variável Renda e Pensão também foram excluídas do processo, pois a variável Renda apresentou um comportamento não linear em relação ao seu valor do Risco Relativo e a variável Pensão possuía poucos clientes com essa característica.

Tabela 2.1: Proposta de crédito da empresa (2004).

Variável	Descrição
Sexo	Feminino ou Masculino
Idade	Idade do cliente no dia do cadastro (anos)
Estado Civil	Casado, solteiro, divorciado, viúvo, etc
Escolaridade	Fundamental, médio, superior completo ou incompleto
Renda	Valor da renda (R$)
Tipo de Renda	Renda declarada ou comprovada
Profissão	Profissão ou cargo do cliente
Tipo Ocupação	Assalariado, autônomo, profissional liberal, etc
CEP Residencial	CEP do local onde reside
CEP Comercial	CEP do local onde trabalha
Tempo Serviço	Tempo no local onde trabalha (meses)
Crédito 3°s	Tem crédito em outros estabelecimentos?
Tipo residência	Própria, alugada, cedida ou com pais
Cidade Nascimento	Cidade de naturalidade do cliente
Filho	Tem filhos?
Pensão	Paga pensão alimentícia?

As variáveis Profissão, CEP Residencial, Cidade de Nascimento e CEP Comercial, possuíam uma grande quantidade de atributos e portanto, foram criados grupos conforme o seu Risco Relativo. Sendo assim, as variáveis foram agrupadas seguindo a seguinte escala: péssimo (RR $<$ 0,50); muito mau (RR entre 0,50 e 0,67); mau (RR entre 0,67 e 0,90); neutro (RR entre 0,90 e 1,10); bom (RR entre 1,10 e 1,50); muito bom (RR entre 1,50 e 2,00) e excelente (RR maior que 2,00) (Selau, 2008). Estabeleceu-se a ocorrência de no mínimo 30 observações em cada atributo.

As variáveis CEP Residencial e Comercial, compostas por oito digítos foram separadas para melhorar a análise da informação. Inicialmente, segregou-se os atributos em dois dígitos, após três dígitos e em seguida quatro dígitos. Determinou-se também a ocorrência de no mínimo 30 casos em cada um. Por exemplo, se o total de casos com o CEP residencial inicial 914 for representativo, analisa-se a divisão de CEP de 9140 a 9149.

Realizou-se o agrupamento das variáveis citadas, criando assim sete grupos, do cliente péssimo ao excelente, baseados pelo seu Risco Relativo. Cada grupo foi transformado em uma variável *dummy* (0 ou 1), as quais serão testadas como variáveis preditoras na construção dos modelos, assim como as demais variáveis originais.

Como discutido anteriormente, após a limpeza da base de dados e agrupamento das variáveis de acordo com o valor do seu Risco Relativo, a amostra possuia 11389 clientes. Os primeiros modelos de Regressão Logística e Regressão Logística Geograficamente Ponderada foram criados com base nessa amostra, porém como não encontrou-se variáveis com informações sobre a latitude e longitude para serem usadas no modelo de RLGP, optou-se pelo uso de apenas uma das informações de CEP dos clientes que foi a do CEP residencial. Após a aplicação da função *buscamulti* do pacote *CepR* não foi possivel encontrar todas as latitudes e longitudes apenas pelo CEP residencial fazendo com que a base de dados da Regressão Logística Geograficamente Ponderada tivesse um número de observações menor. Decidiu-se que para uma comparação justa entre ambas as regressões deveriamos usar um mesmo número de observações. Sendo assim, utilizando apenas os CEP's que começavam

com os dígitos 90 e 91 correspondentes a cidade de Porto Alegre, a modelagem de ambas as regressões foi feita utilizando uma amostra de treino e teste de 2014 e 515 clientes, respectivamente.

As relações dos atributos classificados em cada um dos sete grupos para as variáveis profissão, cidade de nascimento, CEP Residencial e Comercial são apresentados no Apêndice A, Apêndice B, Apêndice C e Apêndice D, respectivamente.

Os resultados da construção dos modelos de previsão de risco de crédito com utilização das técnicas de Regressão Logística e Regressão Logística Geograficamente Ponderada são apresentados na seção Resultados.

Todas as técnicas de modelagem utilizadas neste trabalho foram construídas no *Software* R, versão 4.0.4, e foram aplicadas conforme a lógica descrita na sequência deste texto. A sintaxe utlizada está disponível no Apêndice E deste trabalho.

Para a Regressão Logística, criou-se um modelo com todas as covariáveis disponíveis, por meio da função *glm()*. Por meio do método *stepwise* foi feita a seleção das variáveis *dummies* que fariam parte da construção dos modelos, sendo selecionadas apenas variáveis que demonstraram significância, um total de 19 variáveis preditoras, de um total de 61 *dummies* criadas. A Regressão Logística Geograficamente Ponderada foi implementada pela função *ggwr.basic*, do pacote *GWmodel*, onde desenvolveu-se 2 modelos (um para cada tipo de função de ponderação Gaussiana) e as variáveis preditoras foram as mesmas selecionadas pelo modelo de Regressão Logística.

Após a construção dos modelos a partir da amostra de treino, é necessário analisar o comportamento dos escores na amostra de teste, para verificar a adequação dos modelos encontrados. Este processo se torna importante para que as decisões não sejam viciadas e que os modelos sejam bons apenas para dados aos quais foram treinados. Assim, espera-se que os escores nesta amostra de teste não sejam muito diferentes dos escores encontrados na amostra de treino.

3 Referencial Teórico

O referencial teórico abordou os conceitos de avaliação do risco de crédito, modelos de *Credit Scoring*, modelos de Regressão Logística, Regressão Geograficamente Ponderada e Regressão Logística Geograficamente Ponderada, percentual de acerto, área abaixo da curva ROC, Critério de Informação de Akaike (AIC), e teste KS.

3.1 Avaliação do Risco de Crédito

Conceder crédito é uma das atividades básicas das instituições financeiras, porém no decorrer deste negócio os bancos se expõem a diversos tipos de riscos, o mais comum é o risco de crédito (Ferreira *et al.*, 2012). O simples ato de emprestar dinheiro a alguém traz a possibilidade de que o dinheiro não seja reembolsado e o retorno seja incerto. Esse é o risco de crédito, que pode ser definido como: o risco da contraparte no contrato de concessão de crédito não cumprir sua promessa.

Caouette *et al.* (2000) afirmam que, se crédito é um valor que deverá ser devolvido em um prazo determinado acrescido de juros, o risco de crédito é chance de que isso não aconteça.

Há duas maneiras de avaliar o risco de clientes em potencial: através do julgamento, uma forma mais qualitativa de análise, ou através de modelos de avaliação para classificar os clientes, o que envolve uma análise quantitativa.

As grandes empresas que trabalham com concessão de crédito geralmente utilizam as duas formas combinadas. Na avaliação do risco de crédito são utilizados modelos chamados *Credit Scoring* (pontuação de crédito) para classificar o risco de o tomador de crédito se tornar inadimplente e auxiliar na tomada de decisão de conceder ou não o crédito (Camargos *et al.*, 2012).

3.2 Modelos de Credit Scoring

De acordo com Mileris (2012), a idéia do modelo de *Credit Scoring* é compactar várias informações quantitativas e qualitativas dos clientes em uma pontuação para refletir a capacidade de pagamento.

Dentre as vantagens de se utilizar os modelos de *Credit Scoring*, Caouette *et al.* (2000) enfatizam a objetividade, a consistência e a velocidade, e acreditam que, se desenvolvidas adequadamente, as práticas discriminatórias nos empréstimos podem ser eliminadas.

Na década de 1950, os modelos de crédito se tornaram comuns no setor bancário dos EUA. Os primeiros modelos determinavam certas características com base em pesos pré-determinados, assim somavam-se os pontos e gerava o escore de classificação (Gonçalves *et al.*, 2013).

O crescimento do uso de modelos na década de 1960 mudou os negócios no mercado dos EUA. O alto volume de solicitações de cartão de crédito exigiu dos bancos maior velocidade e automatização nas concessões, e para isso foi usado modelos de *Credit Scoring*. Desde o final da década de 80, o sucesso alcançado por meio dos cartões permitiu que o sistema de pontuação de crédito fosse utilizado pelos bancos para a adesão de outros produtos (Thomas, 2000).

Não apenas as empresas do setor financeiro, mas os grandes varejistas também começaram a usar o modelo de *Credit Scoring* para vender crédito a seus consumidores.

No Brasil, a história é ainda mais curta. Após a implementação do plano Real para alcançar a estabilidade, as instituições financeiras começaram a usar o modelo de *Credit Scoring* em larga escala apenas em meados da década de 1990 (Camargos *et al.*, 2012).

3.3 Modelos de Regressão Logística

3.3.1 Regressão Logística

A Regressão Logística é a técnica mais usada no mercado para o desenvolvimento de modelos de *Credit Scoring* (Crook *et al.*, 2007; Gouvêa, 2015). Diferentemente da análise discriminante, ela não precisa assumir a normalidade das variáveis preditoras e é mais robusta quando a independência das variáveis não é satisfeita (Hair *et al.*, 2009; Gouvêa, 2015).

Em um modelo de Regressão Logística, a variável dependente é uma variável binária e as variáveis preditoras podem ser categóricas (desde que dicotomizadas após transformação) ou contínuas. Considere o caso em que os indivíduos que podem ser classificados como bons ou maus clientes (Gouvêa, 2015).

A variável dependente binária Y pode assumir os valores: 1, se o i-ésimo indivíduo pertence à categoria dos bons e 0 se pertence à categoria dos maus (Tsai, 2010).

Seja, X = $(1, X_1, X_2, ..., X_n)$ um vetor onde o primeiro elemento é igual a 1 (constante) e os demais representam as n variáveis preditoras do modelo.

O modelo de Regressão Logística é um caso particular dos Modelos Lineares Generalizados (Neter *et al.*, 1996), tal que:

$$\beta' X = \ln\left(\frac{p(\mathbf{X})}{1 - p(\mathbf{X})}\right) \tag{3.1}$$

onde:

$\beta' = (\beta_1, \beta_2, \beta_3, ..., \beta_n)$ é o vetor de parâmetros associados às variáveis e p($\mathbf{X}$) = E(Y=1|$\mathbf{X}$) é a probabilidade de o indivíduo ser classificado como bom, dado o vetor X.

Essa probabilidade é expressa por (Neter *et al.*, 1996).

$$p(\mathbf{X}) = E(Y = 1|\mathbf{X}) = E(Y) = \frac{e^{\beta' X}}{1 + e^{\beta' X}} \tag{3.2}$$

3.3.2 Regressão Geograficamente Ponderada

A técnica de Regressão Geograficamente Ponderada (RGP); proposta por Fotheringham *et al.* (2002), utilizada para modelar processos heterogêneos espacialmente e é um método que procura fornecer uma versão local da análise de regressão linear. Com o auxílio de subamostras de observações ponderadas pela distância geográfica, obtem-se coeficientes locais para cada ponto no espaço. O modelo gera uma sequência de regressões lineares, estimadas para cada região, usando subamostras dos dados, ponderadas pela distância (Almeida, 2012).

A ideia é ajustar um modelo de regressão a cada ponto observado, ponderando todas as demais observações como função da distância a este ponto. Se existe alguma variação geográfica na relação, então essa variação fica incluída como erro.

Dado um modelo de regressão linear, a expressão equivalente para a RGP é:

$$y_i = \beta_0(u_i, v_i) + \sum_k \beta_k(u_i, v_i)x_{ik} + \varepsilon_i \tag{3.3}$$

(u_i, v_i) representa as coordenadas do i-ésimo ponto no espaço, para i = 1,...,n $\beta_k(u_i, v_i)$ é o parâmetro para a k-ésima variável explicativa, em função do local da i-ésima observação, x_{ik} é o valor da k-ésima variável explicativa para o local i.

O termo de erro aleatório segue distribuição normal com média zero e variância constante.

No modelo RGP, assume-se que as informações mais próximas do ponto de regressão têm maior probabilidade de influenciá-lo. Tal ponderação é feita pela função Kernel espacial. O Kernel usa distância (d_{ij}) entre dois pontos geográficos representando duas regiões e um parâmetro da largura da banda (b) para determinar um peso (w_{ij}) entre essas duas regiões, que é inversamente relacionado à distância geográfica (Marconato, 2020). Deve-se especificar (w_{ij}) como uma função contínua de (d_{ij}), que representa a distância entre i e j.

Os pressupostos do modelo clássico de regressão linear permanecem para a RGP. A seguir está a forma matricial para estimação dos parâmetros da RGP:

$$\hat{\beta}(i) = (X'W(u_i, v_i)X)^{-1}X'W(u_i, v_i)y \tag{3.4}$$

onde:

$$W(u_i, v_i) = \begin{bmatrix} W_{i1} & 0 & \cdots & 0 \\ 0 & W_{i2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & W_{in} \end{bmatrix} \tag{3.5}$$

Se substituirmos todos os pesos w_{ij} por 1 teremos a matriz identidade que se substituída em (3.4) retorna ao modelo clássico de regressão linear.

As duas principais funções de ponderação espacial encontradas na literatura são a função Gaussiana e a função Biquadrática (Fotheringham *et al.*, 2002).

Funções de Ponderação	Fórmula das Funções de Ponderação
Gaussiana Fixa	$w_{ij} = exp\left\{-\frac{1}{2}\left(\frac{d_{ij}}{b}\right)^2\right\}$
Biquadrática Fixa	$w_{ij} = \left[1-\left(\frac{d_{ij}}{b}\right)^2\right]^2$ $se\ d_{ij} < b, e\ w_{ij} = 0\ caso\ contrário$
Gaussiana Variável	$w_{ij} = \exp\left\{-\frac{1}{2}\left(\frac{d_{ij}}{b_{i(k)}}\right)^2\right\}$
Biquadrática Variável	$w_{ij} = \left[1-\left(\frac{d_{ij}}{b_{i(k)}}\right)^2\right]^2$ $se\ d_{ij} < b_{i(k)}, e\ w_{ij} = 0\ caso\ contrário$

Figura 3.1: Funções de Ponderação.
Fonte: Fotheringham *et al.* (2002).

Nota-se, pela Figura 3.1, que existem dois tipos de expressões para cada uma das funções Gaussiana e Biquadrática, que se diferenciam por meio da escolha do parâmetro b *(bandwidth)* a ser utilizado (se fixo ou variável). O parâmetro d_{ij} contido nas funções de ponderação representa a distância do ponto i ao ponto j, o parâmetro b é o *bandwidth* (parâmetro de suavização) fixo e o parâmetro $b_i(k)$ representa o *bandwidth* variável, sendo que a letra k representa o número de vizinhos mais próximos do ponto i. Essas funções fazem com que o peso diminua à medida que a distância aumente (Albuquerque & Silva, 2017).

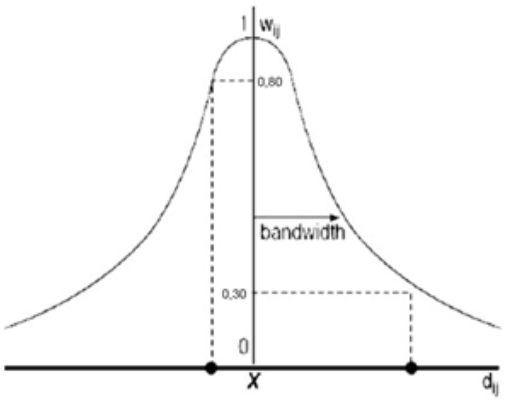

Figura 3.2: *Bandwidth* ou Parâmetro de Suavização.
Fonte: Adaptado de (Fotheringham *et al.*, 2002).

Ao desenvolver um modelo utilizando o *bandwidth* fixo, por meio da RGP ele deve ser especificado por seu valor em unidade de distância; no entanto, ao utilizar o *bandwidth* variável, deve-se definir k vizinhos mais próximos (fixos) a serem usados no modelo e, com base nessa quantidade k, o valor do *bandwidth* varia entre as regiões do estudo (Albuquerque & Silva, 2017).

3.3.3 Regressão Logística Geograficamente Ponderada

Quando a variável de interesse é do tipo binária, a aplicação da RGP deve ser feita por meio da Regressão Logística Geograficamente Ponderada, cuja equação é

usada para obtenção da probabilidade de ocorrência do evento de interesse, conforme a equação (3.6)

$$\ln(\frac{\pi(x_j)}{1-\pi(x_j)}) = \beta_0(u_i, v_i) + \sum_{k=1}^{p} \beta_k(u_i, v_i)x_{jk} + \varepsilon_i \tag{3.6}$$

$\pi(x_j)$ é a probabilidade do j-ésimo cliente se tornar inadimplente.

A função $\beta_k(u_i, v_i)$ representa os parâmetros (coeficientes) das k variáveis do modelo, que variam de acordo com a região i de coordenadas latitude e longitude (u_i, v_i) (Albuquerque & Silva, 2017).

A estimação dos parâmetros é calculada através do método da máxima verossimilhança, sendo a função de verossimilhança representada pela equação (3.7)

$$L(\beta(u_i, v_i)) = (\prod_{j=1}^{n}[1 + exp(\sum_{k=0}^{p} \beta_k u_i, v_i x_j k)]^{-1})exp[\sum_{k=0}^{p}(\sum_{j=1}^{n} y_j x_j k)\beta_k(u_i, v_i)] \tag{3.7}$$

A matriz W(u_i, v_i), descrita conforme a equação 3.3.2, possui um peso w_{ij} em seus elementos (calculado pela função de ponderação conforme a Figura 3.1) e é usada para ponderar geograficamente as observações de cada conjunto na estimação de parâmetros $\beta_k(u_i, v_i)$, ou seja, a matriz é responsável por atribuir maior peso às observações geograficamente mais próximas da região i ao estimar seus parâmetros e atribuir pesos menores ou zeros às observações mais distantes (dependendo da função de ponderação selecionada). A matriz $W(u_i, v_i)$ também altera de acordo com a localização de cada tomador de crédito e forma a função de verossimilhança, conforme a equação (3.8)

$$\ln[L^*(\beta(u_i, v_i))] = \sum_{k=0}^{p}(\sum_{j=1}^{n} w_j(u_i, v_i)y_j x_j k)\beta_k(u_i, v_i) - \sum_{j=1}^{n} w_j(u_i, v_i) ln[1+exp(\sum_{k=0}^{p} \beta_k(u_i, v_i)x_{jk})] \tag{3.8}$$

Semelhante ao modelo de regressão logística, depois que (3.7) é diferenciado em função de $\beta(u_i, v_i)$ e igualado a zero, método dos mínimos quadrados reponderados iterativos (MQRI) são usados para estimar a iteração de parâmetros do modelo. Cabe ressaltar que esse processo de maximização é realizado para cada função relacionada a cada área i em estudo (Albuquerque & Silva, 2017).

3.4 Avaliação dos Modelos de Regressão Logística

3.4.1 Matriz de Confusão

A matriz de confusão será utilizada para avaliar a acurácia dos modelos. É uma tabela de contingência em que nas linhas estão os valores previstos e nas colunas os valores observados. A Figura 3.3 ilustra a matriz de confusão.

Valor Predito	Valor Observado	
	Sim	Não
Sim	Verdadeiro positivo (VP)	Falso positivo (FP)
Não	Falso negativo (FN)	Verdadeiro negativo (VN)

Figura 3.3: Matriz de Confusão.
Fonte: Crook *et al.* (2007).

- VP: Verdadeiro Positivo - quantidade de clientes bons classificados como bons
- VN: Verdadeiro Negativo - quantidade de clientes maus classificados como maus
- FP: Falso Positivo - quantidade de clientes maus classificados como bons
- FN: Falso Negativo - quantidade de clientes bons classificados como maus

Existem dois tipos de erro que um modelo classificador pode cometer: reprovar clientes que deveriam ser aprovados (Falso Negativo - FN) ou aprovar clientes que deveriam ser reprovados (Falso Positivo - FP), sendo que este último, também conhecido como Erro do tipo II, é considerado o pior dos dois erros, pois esse cliente seria aprovado e poderia gerar prejuízos financeiros para a instituição que precisariam em média de 5 clientes bons aprovados para compensá-lo. Portanto, é importante prever e reduzir a inadimplência, pois os prejuízos com créditos mal sucedidos deverão ser cobertos com a cobrança de altas taxas de juros em novas concessões (Selau, 2008).

A acurácia do modelo é calculada pela proporção de VP e VN, ou seja, é a proporção de casos que foram corretamente previstos, sejam eles verdadeiro positivo ou verdadeiro negativo, conforme a equação (3.9)

$$ACC = \frac{VP + VN}{VP + VN + FP + FN} \tag{3.9}$$

A sensibilidade e a especificidade podem ser calculadas a partir da figura 3.3, respectivamente equações (3.10) e (3.11)

$$\frac{VP}{VP + FN} \tag{3.10}$$

$$\frac{VN}{VN + FN} \tag{3.11}$$

3.4.2 Curva ROC e AUC

A avaliação do desempenho do modelo é uma tarefa muito importante. Na área de risco de crédito uma das ténicas mais utilizadas para auxiliar na classificação de clientes como bons ou maus pagadores é a curva ROC, a qual obtemos esboçando um gráfico da taxa de verdadeiro positivo (sensibilidade) contra a taxa de falso positivo. Segundo Hosmer & Lemeshow (2000) a regra geral para avaliação do resultado da área abaixo da curva ROC de modelos de *Credit Scoring* é dada por:

$$
\begin{aligned}
\text{área} < 0,7 &\rightarrow \text{baixa discriminação} \\
0,7 \leq \text{área} < 0,8 &\rightarrow \text{discriminação aceitável} \\
0,8 \leq \text{área} < 0,9 &\rightarrow \text{discriminação excelente} \\
\text{área} > 0,9 &\rightarrow \text{discriminação excepcional}
\end{aligned}
$$

Figura 3.4: Regra geral da curva ROC.
Fonte: Hosmer & Lemeshow (2000).

Modelos com poder de discriminação excepcional concentram-se no canto superior esquerdo da curva ROC, pois conforme a sensibilidade aumenta há pouca perda de especificidade. Porém, modelos com baixa discriminação aproximam-se da diagonal.

Assim, quanto maior o AUC, melhor o modelo consegue prever bons pagadores e maus pagadores de forma correta. Um modelo excelente tem AUC próximo ao 1, o que significa que ele classifica as previsões corretamente. Um modelo ruim tem AUC próximo do 0, o que significa que ele classifica as previsões incorretamente, ou seja, está prevendo um bom pagador como mau e um mau pagador como bom. E quando a AUC é 0,5, significa que o modelo é incapaz de separar as classes.

3.4.3 Critério de Informação de Akaike

Um critério de seleção que tem sido muito utilizado em seleção de modelos é o critério de Akaike (AIC), conforme a equação (3.12)

$$AIC = -2\sum_{i=1}^{n} \ln L(\hat{\mu}_i, y_i) + 2 * (\text{número de parâmetros}) \tag{3.12}$$

y_i é o i-ésimo valor da resposta e $\hat{\mu}_i$ é a estimativa de y_i.

O AIC torna-se útil quando são comparados diversos modelos. O melhor modelo será aquele que apresentar o menor valor de AIC.

O AIC foi desenvolvido a partir da divergência de Kullback-Leibler (K-L), que mesura a diferença entre dois modelos. É recomendada a utilização do AIC apenas quando a razão entre o número de observações e variáveis preditoras (n/p) for maior ou igual a 40 (Burnham, 2002). A expressão AIC pode ser simplificada, conforme a equação (3.13)

$$AIC = n \ln(\hat{\sigma}_p^2) + 2 * (p + 1) \tag{3.13}$$

$\hat{\sigma}_p^2$ é o estimador de máxima verossimilhança da variância do erro, conforme a equação (3.14).

$$\hat{\sigma}_p^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2}{n} \tag{3.14}$$

3.4.4 Teste de Kolmogorov-Smirnov

O valor do teste de Kolmogorov-Smirnov (KS); segundo Crook *et al.* (2007), é uma importante medida de separação, muito utilizada na prática (Picinini *et al.*,

2003). Trata-se de um teste não paramétrico para determinar se duas amostras foram extraídas da mesma população (ou de populações com distribuições similares) (Siegel, 1975).

Em ambas as populações (bons e maus clientes) os intervalos são divididos em tamanhos iguais de pontuação (escores) e determina-se a frequência acumulada de cada um. Após é calculado a diferença entre cada frequência acumulada e o valor do teste é o resultado da maior diferença entre as frequências acumuladas.

Quanto à definição de um valor ideal para o teste KS em modelos de *Credit Scoring*, Picinini *et al.* (2003) sugere: "O teste de Kolmogorov-Smirnov (KS) é utilizado no mercado financeiro como um dos indicadores de eficiência de modelos de *Credit Scoring*, sendo que o mercado considera um bom modelo aquele que apresente um valor de KS igual ou superior a 30". Segundo Picinini *et al.* (2003), "modelos de *Credit Scoring* com taxas de acerto acima de 65% (obtido pela matriz de confusão) são considerados bons por especialistas".

4 Resultados

Nesta seção são apresentados os resultados dos dois modelos criados conforme cada abordagem utilizada. Por fim, será feita uma discussão sobre os resultados encontrados.

4.1 Regressão Logística

O modelo foi desenvolvido utilizando a amostra de treino com 2014 observações. As variáveis escolhidas para compor o modelo foram as *dummies* selecionadas pelo método *stepwise*. Criou-se um modelo com todas as variáveis significativas (p-valor abaixo de 0,05) a fim de compará-lo com a técnica de Regressão Logística Geograficamente Ponderada. O modelo final possui 19 variáveis preditoras, as variáveis com sinais positivos revelam associações com ser bom pagador e as de sinais negativos com ser mau pagador. Os indicadores de desempenho do modelo na capacidade de predição foram avaliados na amostra de teste e são apresentados na Tabela (4.1). A lista de variáveis *dummies* são apresentadas na Tabela (4.2).

Tabela 4.1: Matriz de Confusão da amostra de teste - Regressão Logística.

	Valor Observado		
Valor Predito	Bom	Mau	Total
Bom	103	70	173
Mau	102	240	342
Total	205	310	515

Portanto, dos 515 dados contidos na amostra de teste, o modelo previu corretamente que 240 são inadimplentes e 103 são adimplentes, o que resulta numa precisão de 66,60%, indicador o qual mantém-se bem próximo do encontrado na amostra de treino (66,88%). A Tabela (4.3) apresenta as medidas referentes à AUC e ao teste KS, o valor do AUC na amostra de treino foi de 73,08% e teve uma leve queda na amostra de teste, com o valor de 72,23%. A capacidade de diferenciação da curva de bons pagadores para a de maus é apresentada pelo teste KS, com os valores de 34,17% e 32,73% para as amostras de treino e teste, respectivamente.

Tabela 4.2: Variáveis do Modelo via Regressão Logística e seus coeficientes.

Variáveis	Coeficientes	Intervalos de Confiança
Intercepto	1.1119	0.6982 ; 1.5312
DIDAD1	-0.5567	-0.9430 ; -0.1760
DIDAD2	-0.4469	-0.7710 ; -0.1247
DIDAD7	0.5308	0.2197 ; 0.8484
DIDAD8	1.2722	0.8857 ; 1.6741
DSEXOF	0.3877	0.1696 ; 0.6058
DPRIM	-0.4671	-0.6802 ; -0.2552
DCASADO	0.5335	0.2963 ; 0.7740
DTSERV2	-0.8064	-1.4408 ; -0.2012
DTSERV8	1.1521	0.1464 ; 2.4104
DTSERV9	0.8792	0.3212 ; 1.4990
DOCUP_AS	-0.3770	-0.6819 ; -0.0744
DOCUP_AU	-0.4200	-0.7210 ; -0.1205
DRES_ALU	-0.6855	-1.0314 ; -0.3399
DRES_OUT	-0.4171	0.0310 ; 0.8110
DGCEPRE1	-1.1905	-2.1223 ; -0.2829
DGCEPRE2	-0.4367	-0.7286 ; -0.1469
DGCEPRE3	-0.4135	-0.6762 ; -0.1544
DCIDNA2	-0.4388	-0.6601 ; -0.2188
DCIDNA7	0.9369	0.2014 ; 1.7968

Tabela 4.3: Indicadores de desempenho - Modelo via Regressão Logística.

Medida	Treino	Teste
AUC	73,08%	72,23%
KS	34,17%	32,73%

O valor encontrado para o Critério de Informação de Akaike do modelo de Regressão Logística foi de 2339,31. Sendo esse o valor para comparação do modelo via Regressão Logística Geograficamente Ponderada, os resultados são apresentados a seguir.

4.2 Regressão Logística Geograficamente Ponderada

Foram densenvolvidos dois modelos utilizando a técnica de Regressão Logística Geograficamente Ponderada, sendo um para cada função de ponderação Gaussiana conforme Figura 3.1. As variáveis preditoras utilizadas foram as mesmas do modelo de Regressão Logística, o melhor modelo via Regressão Logística Geograficamente Ponderada, segundo o critério AIC, foi o modelo Gaussiano Variável, com valor de 1812 vizinhos mais próximos para estimar os *bandwidths* variáveis.

A Tabela 4.4 contém os coeficientes estimados pelo modelo Regressão Logística Geograficamente Ponderada, algumas das médias dos coeficientes ficaram bem próximas dos coeficientes do modelo Regressão Logística apresentados na Tabela (4.2).

Tabela 4.4: Coeficientes estimados do modelo RLGP Gaussiano Variável.

Variáveis	Coeficientes
Intercepto	1.0620
DIDAD1	-0.5834
DIDAD2	-0.4462
DIDAD7	0.5174
DIDAD8	1.3174
DSEXOF	0.3974
DPRIM	-0.4708
DCASADO	0.5402
DTSERV2	-0.8324
DTSERV8	1.3198
DTSERV9	0.9033
DOCUP_AS	-0.3267
DOCUP_AU	-0.3800
DRES_ALU	-0.6499
DRES_OUT	-0.4338
DGCEPRE1	-1.2215
DGCEPRE2	-0.4592
DGCEPRE3	-0.4118
DCIDNA2	-0.4187
DCIDNA7	0.8782

Os indicadores de desempenho do modelo na capacidade de predição foram avaliados na amostra de teste e são apresentados na Tabela 4.5.

Tabela 4.5: Matriz de Confusão da amostra de teste - RLGP.

	Valor Observado		
Valor Predito	Bom	Mau	Total
Bom	109	64	173
Mau	107	235	342
Total	216	299	515

Portanto, dos 515 dados contidos na amostra de teste, o modelo previu corretamente que 235 são inadimplentes e 109 são adimplentes, o que resulta numa precisão de 66,79%, indicador o qual mantém-se bem próximo do encontrado na amostra de treino (67,77%). A Tabela (4.6) apresenta os demais indicadores utilizados para fins de comparação de desempenho, onde nota-se que a precisão do modelo praticamente manteve-se estável nas duas amostras, já a capacidade de diferenciação de bons pagadores para os maus diminuiu cerca de 4 pontos percentuais na amostra de teste quando comparada com a de treino.

Tabela 4.6: Indicadores de desempenho - Modelo via RLGP.

Medida	Treino	Teste
AUC	74,22%	72,83%
KS	36,64%	32,75%

4.3 Discussão

A Tabela 4.7 apresenta as comparações dos indicadores de desempenho dos dois modelos abordados neste trabalho, onde se nota pequena diferença entre os valores dos indicadores dos dois modelos.

Tabela 4.7: Comparação dos indicadores de desempenho.

	Treino			Teste		
Modelo	% Acerto	AUC	KS	% Acerto	AUC	KS
RL	66,88%	73,08%	34,17%	66,60%	72,23%	32,73%
RLGP	67,77%	74,22%	36,64%	66,79%	72,83%	32,75%

Modelo	AIC
RL	2339,31
RLGP	2336,68

Pela tabela é possível notar que na amostra de treino o modelo que obteve o maior percentual de acerto foi o da Regressão Logística Geograficamente Ponderada e o mesmo acontece no conjunto de teste, o que corrobora com Picinini *et al.* (2003), "modelos de *Credit Scoring* com taxas de acerto acima de 65% (obtido pela matriz de confusão) são considerados bons por especialistas".

O modelo de Regressão Logística Geograficamente Ponderada foi o modelo que apresentou um menor valor de AIC e uma maior acurácia, que indica um melhor percentual de acertos e menor percentual de Falsos Positivos.

Como discutido anteriormente, a AUC é um critério que mede a discriminação entre as classes estudadas, neste caso, bons e maus clientes. Na amostra de teste, a Regressão Logística Geograficamente Ponderada apresentou desempenho superior à Regressão Logística e no conjunto de treino, este resultado se manteve, porém com uma diferença maior.

Em ambos os conjuntos de amostras utilizados os dois modelos apresentaram um valor de KS considerado bom. O modelo que apresentou melhor desempenho em relação ao KS foi a Regressão Logística Geograficamente Ponderada nos dois conjuntos amostrais testados, porém na amostra de teste a diferença deste valor foi muito pequena.

5 Considerações Finais

Este trabalho fez o uso de um banco de dados real utilizado na concessão de crédito, assim foi possível o desenvolvimento de modelos de *Credit Scoring* através de duas metodologias distintas: Regressão Logística e Regressão Logística Geograficamente Ponderada.

Este trabalho teve como objetivo principal comparar o desempenho das duas metologias citadas anteriormente. A metodologia Regressão Logística é bastante difundida no setor financeiro, sendo utilizada neste estudo para desenvolver um modelo de *Credit Scoring* para a cidade de Porto Alegre. A metodologia Regressão Logística Geograficamente Ponderada, por outro lado, ainda é pouco conhecida. O modelo considera o fator geográfico do cliente como o peso na estimativa dos parâmetros, atribuindo pesos diferentes para cada localização.

Os resultados dos indicadores usados para comparar os desempenhos dos dois modelos desenvolvidos se mostraram bem próximos e, podemos considerar que ambos são semelhantes em termos de capacidade de previsão de perdas financeiras para o banco de dados analisado. O modelo de Regressão Logística Geograficamente Ponderada se mostrou melhor, obtendo um desempenho superior em todas as métricas utilizadas: percentual de acerto, área abaixo da curva ROC, AIC e medida do teste KS.

Com relação às limitações do estudo, a dificuldade para encontrar as coordenadas de latitude e longitude a partir apenas dos dados dos CEP's residenciais como informação do endereço, limitou a aplicação da técnica de Regressão Logística Geograficamente Ponderada à cidade de Porto Alegre. Os resultados encontrados na comparação dos modelos podem não terem sido tão expressivos por conta da variável geográfica. Acredita-se que a perda de informação gerada pela conversão do CEP em latitude e longitude prejudicou o desempenho do modelo por fazer com que o número de observações disponíveis diminuísse.

Como tópicos de pesquisas futuras, é possível aplicar o método Regressão Logística Geograficamente Ponderada para públicos-alvo de outras regiões geográficas e assim desenvolver modelos de *Credit Scoring* utilizando variáveis preditoras diferentes, utilizar uma base de dados maior ou então aplicar essa metodologia utilizando outra função de ponderação para analisar se há uma melhora no desempenho do modelo.

Referências Bibliográficas

Albuquerque, P. H. M.and Medina, F. A. S, & Silva, A. R. 2017. Regressão Logística Geograficamente Ponderada Aplicada a Modelos de Credit Scoring. *Revista de Contabilidade e Finanças*, **28**(73), 93–112.

Almeida, E. 2012. *Econometria espacial aplicada*. first edn. Alinea.

Atkinson, P. M., German, S. E., Sear, D. A., & Clark, M. J. 2003. Exploring the relations between riverbank erosion and geomorphological controls using geographically weighted logistic regression . *Geographical Analysis*, **27**(2), 58–82.

Burnham, K.P. Anderson, D.R. 2002. *Model selection and multimodel inference. A practical information - theoretic approach*. first edn. Springer.

Camargos, M. A., Camargos, M. C. S., & Araújo, E. A. T. 2012. A inadimplência em um programa de crédito de uma instituição financeira pública de Minas Gerais: uma análise utilizando regressão logística. *REGE*, **3**(2), 473–492.

Caouette, J., Altmano, E., & Narayanan, P. 2000. *Gestão do Risco de Crédito*. second edn. Qualitymark.

Crook, J. N., Edelman, D. B., & Thomas, L. C. 2007. Recent developments in consumer credit risk assessment. *European Journal of Operational Research*, **183**(3), 1447–1465.

Fernandes, G. B., & Artes, R. 2016. Spatial dependence in credit risk and its improvement in credit scoring. *European Journal of Operational Research*, **249**(2), 517–524.

Ferreira, M. A. M., Celso, A. S. S., & Barbosa Neto, J. E. 2012. Aplicação do modelo logit binomial na análise do risco de crédito em uma instituição bancária. *Revista de Negócios*, **17**(1), 41–59.

Fotheringham, A. S., Brunsdon, C., & Charlton, M. 2002. *Geographically weighted regression: the analysis of spatially varying relationships*. first edn. John Wiley Sons.

Gonçalves, E. B., Gouvêa, M. A., & Mantovani, D. M. N. 2013. Análise de risco de crédito com aplicação de regressão logística e redes neurais. *Revista Contabilidade Vista Revista*, **24**(4), 96–123.

Gouvêa, M. A.and Gonçalves, E. B.and Mantovani D. M. N. 2015. Análise de Risco de Crédito com Aplicação de Regressão Logística e Redes Neurais. *Contabilidade Vista Revista*, **24**(4), 96–123.

Hair, JR. J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. 2009. *Análise multivariada de dados*. first edn. Bookman.

Hand, D. J., & Henley, W. E. 1997. Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society*, **160**, 523–541.

Hosmer, D. W., & Lemeshow, S. 2000. *Applied logistic regression*. first edn. John Wiley Sons.

Lessmann, S., Baesens, B., V., Seow H., & Thomas, L. C. 2015. Benchmarking state of the art classification algorithms for credit scoring: An update for research. *European Journal of Operational Research*, **247**, 124–136.

Marconato, M., Parré J. L. Coelho M. H. 2020. Análise fiscal dos municípios brasileiros no ano de 2016, a partir do modelo RPG. *Revista de Economia Mackenzie*, **17**(1), 12–41.

Mileris, R. 2012. Macroeconomic Determinants of Loan Portfolio Credit Risk in Banks. *Inzinerine Ekonomika-Engineering Economics*, **23**(5), 496–504.

Neter, J., Kutner, M. H., Nachtshein, C. J., & Wasserman, W. 1996. *Applied linear statistical models*. first edn. Irwin.

Picinini, R., Oliveira, G. M. B., & Monteiro, L. H. A. 2003. Mineração de critério de credit scoring utilizando algoritmos genéticos. *Simpósio Brasileiro de Automação Inteligente*, **6**(5).

Selau, Lisiane Priscila Roldão. 2008. *Construção de modelos de previsão de risco de crédito*. M.Phil. thesis, Escola de Engenharia da Universidade Federal do Rio Grande do Sul.

Sicsú, A.L. 2010. *Credit Scoring: desenvolvimento, implantação, acompanhamento*. first edn. Blucher.

Siegel, S. 1975. *Estatística não-paramétrica para as ciências do comportamento*. first edn. McGraw-Hill.

Silva, P. R. 2011. *Psicologia do risco de crédito: análise da contribuição de variáveis psicológicas em modelos de credit scoring*. Ph.D. thesis, Faculdade de Economia, Administração e Contabilidade da Universidade de São Paulo.

Stine, R. (ed). 2011. *Spatial temporal models for retail credit*.

Thomas, L. 2000. A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting*, **16**(2), 149–172.

Tsai, B. 2010. Comparison of Binary Logit Model and Multinomial Logit Model in Predicting Corporate Failure. *Review of Economics Finance*, **1994**, 99–111.

Apêndice

A Agrupamento de profissões

Péssimo Desempenho	BABA	PROMOTOR VENDAS
	COZINHEIRO	ALMOXARIFE
	PINTOR	
Muito Mau Desempenho	AUX PRODUCAO	PEDREIRO
	CABELEIREIRO	PORTEIRO
	CONFEITEIRO	RECEPCIONISTA
	GERENTE	VENDEDOR
	PADEIRO	
Mau Desempenho	AUTONOMO	MANICURE
	AUX ADMINISTRATIVO	MECANICO
	AUX COZINHA	TEC ENFERMAGEM
	AUX SERVICOS GERAIS	VIGILANTE
	COMERCIANTE	
Desempenho Neutro	ATENDENTE	INDUSTRIARIO
	COMERCIARIO	MOTORISTA
	DOMESTICA	
Bom Desempenho	CAIXA	SECRETARIA
	DO LAR	SERVENTE
	PENSIONISTA	
Muito Bom Desempenho	AGRICULTOR	OPERADOR
	BALCONISTA	METALUGICO
	COSTUREIRO	AUX ENFERMAGEM
	DIARISTA	
Excelente Desempenho	APOSENTADO	PROFESSOR

B Agrupamento de cidades de nascimento

Péssimo Desempenho	ALVORADA	
Muito Mau Desempenho	CRUZ ALTA	RIO GRANDE
	ESTEIO	TRAMANDAI
	PORTO ALEGRE	
Mau Desempenho	CANOAS	SAO BORJA
	GRAVATAI	SAO GABRIEL
	IJUI	SAPIRANGA
	NOVO HAMBURGO	SAPUCAIA DO SUL
	PELOTAS	URUGUAIANA
Desempenho Neutro	ALEGRETE	SANTO ANGELO
	CAMAQUA	SAO FRANCISCO DE PAULA
	CANELA	SAO LOURENCO DO SUL
	SANTANA DO LIVRAMENTO	VIAMAO
Bom Desempenho	BAGE	PASSO FUNDO
	BUTIA	RIO PARDO
	CACAPAVA DO SUL	SANTA CRUZ DO SUL
	GUAIBA	SAO JERONIMO
	MONTENEGRO	SAO LEOPOLDO
	OSORIO	SAO LUIZ GONZAGA
Muito Bom Desempenho	CACHOEIRA DO SUL	SANTA ROSA
	CAXIAS DO SUL	SANTA VITORIA DO PALMAR
	PALMEIRA DAS MISSOES	TAQUARA
	SANTA MARIA	
Excelente Desempenho	CANGUCU	SAO SEPE
	ENCRUZILHADA DO SUL	TAPES
	GIRUA	TORRES
	HORIZONTINA	TRES DE MAIO
	ROLANTE	TRIUNFO
	SANTO ANTONIO DA PATRULHA	

C Agrupamento do CEP residencial

	2 PRIMEIRAS POSIÇÕES	3 PRIMEIRAS POSIÇÕES	4 PRIMEIRAS POSIÇÕES
Péssimo Desempenho			**9670** SÃO JERÔNIMO
Muito Mau Desempenho			**9175** Aberta dos Morros - POA **9179** Restinga - POA **9191** Camaquã - POA **9192** Cavalhada/Camaquã - POA **9440** Águas Claras - VIAM **9449** NS Aparecida/Pq Índio Jari - VIAM **9481** Formozo/Passo Feijó - ALVO **9493** Dist Industrial/Cohab - CACH
Mau Desempenho		**902** Farrapos/Navegantes/Humaitá - POA **906** Partenon/Jardim Botânico - POA **912** Protásio Alves/Rubem Berta - POA **915** Lomba do Pinheiro/Agronomia - POA **923** Mathias Velho/Harmonia - CANO **934** Lomba Grande/Santo Afonso - NH **941** GRAVATAÍ **945** Vila Augusta/Jd Universit. - VIAM	**9117** Rubem Berta - POA **9172** Nonoai/Teresópolis - POA **9174** Cavalhada/Vila Nova - POA **9190** Tristeza/Vila Assunção - POA **9326** Centro/Vila Teópolis - EST **9329** Pq Primavera/Pq St Inácio - EST **9353** São Jorge/Vila Diehl - NH **9400** GRAVATAÍ **9441** Centro/Tarumã - VIAM **9442** Vila Elsa/Estalagem - VIAM **9444** Jd Krahe/St Onofre - VIAM **9482** Maria Regina/Sumaré - ALVO **9483** Tijuca/Piratini - ALVO **9485** Aparecida/Jd Algarve - ALVO **9490** Jardim América/Vila City - CACH **9607** Porto/Três Vendas - PEL **9618** CAMAQUÃ **9750** URUGUAIANA
Desempenho Neutro		**908** Santa Tereza/Medianeira - POA **913** Vila Jardim/Vila Ipiranga - POA **914** Protásio Alves/Jardim Carvalho - POA	**9332** Industrial/Ouro Branco - NH **9445** São Lucas/Florescente - VIAM **9447** St Cecília/Viamópolis - VIAM **9480** Maringá/Sumaré - ALVO **9494** Vila Vista Alegre - CACH **9495** Vila Bom Princípio/Pq Matriz - CACH
Bom Desempenho		**922** Fátima/Rio Branco - CANO **924** Igara/São José/Guajuviras - CANO **925** GUAÍBA **930** SÃO LEOPOLDO **938** NOVA HARTZ, SAPIRANGA **955** OSÓRIO, CAPÃO DA CANOA **956** TAQUARA, CANELA, GRAMADO **957** BENTO GONÇALVES, GARIBALDI	**9178** Lami/Belém Novo - POA **9328** Vila Esperança/Pq Amador - EST **9330** Centro - NH **9333** Liberdade/Ideal - NH **9334** Primavera/Petrópolis - NH **9354** Canudos/Mauá - NH **9443** Jardim Krahe/Sítio S.José - VIAM **9496** Pq Granja Esperança - CACH **9601** Centro - PEL **9617** SÃO LOURENÇO DO SUL **9674** ARROIO RATOS e CHARQUEADAS **9754** ALEGRETE
Muito Bom Desempenho	**99** PASSO FUNDO	**950** CAXIAS DO SUL **962** RIO GRANDE, STA VITÓRIA PALMAR **965** CACHOEIRA DO SUL, CAÇAPAVA **970** SANTA MARIA	**9407** GRAVATAÍ
Excelente Desempenho	**98** CRUZ ALTA	**937** CAMPO BOM **958** ESTRELA, TAQUARI, VENÂNCIO AIRES **964** BAGÉ, DOM PEDRITO **966** RIO PARDO, PÂNTANO GRANDE **968** SANTA CRUZ DO SUL **971** SANTA MARIA, ITAARA **973** SÃO GABRIEL, LAVRAS DO SUL	**9352** Guarani/Vila Nova - NH

D Agrupamento do CEP comercial

	2 PRIMEIRAS POSIÇÕES	3 PRIMEIRAS POSIÇÕES	4 PRIMEIRAS POSIÇÕES
Péssimo Desempenho		**912** Protásio Alves/Rubem Berta - POA	**9670** SÃO JERÔNIMO
Muito Mau Desempenho		**900** Centro/Farroupilha/Bom Fim - POA **906** Partenon/Jardim Botânico - POA **932** ESTEIO, SAPUCAIA DO SUL **948** ALVORADA	**9174** Cavalhada/Vila Nova - POA **9190** Tristeza/Vila Assunção - POA **9192** Cavalhada/Camaquã - POA
Mau Desempenho		**901** Azenha/Menino Deus/Praia Belas - POA **902** Farrapos/Navegantes/Humaitá - POA **908** Santa Tereza/Medianeira - POA **910** Passo D'Areia/Jardim Lindóia - POA **911** Sarandi/Rubem Berta - POA **913** Vila Jardim/Vila Ipiranga - POA **915** Lomba do Pinheiro/Agronomia - POA **933** Rio Branco/Primavera/Industrial - NH **940** GRAVATAÍ **949** CACHOEIRINHA **961** CAMAQUÃ, CAPÃO DO LEÃO	**9175** Aberta dos Morros - POA **9179** Restinga - POA **9191** Camaquã - POA **9351** Centro/Hamburgo Velho - NH **9353** São Jorge/Vila Diehl - NH **9602** Fragata/Três Vendas - PEL
Desempenho Neutro		**904** Auxiliadora/Petrópolis - POA	**9380** SAPIRANGA
Bom Desempenho		**905** São João/Floresta/Higienópolis - POA **925** GUAÍBA **934** Lomba Grande/Santo Afonso - NH **955** OSÓRIO, CAPÃO DA CANOA **956** TAQUARA, CANELA, GRAMADO **957** BENTO GONÇALVES, GARIBALDI	**9178** Lami/Belém Novo - POA **9389** NOVA HARTZ **9441** Centro/Tarumã - VIAM **9601** Centro - PEL **9674** ARROIO RATOS e CHARQUEADAS
Muito Bom Desempenho	**99** PASSO FUNDO	**959** LAJEADO, ENCANTADO, PROGRESSO **962** RIO GRANDE, STA VITÓRIA PALMAR	
Excelente Desempenho	**97** SANTA MARIA **98** CRUZ ALTA	**937** CAMPO BOM **950** CAXIAS DO SUL **958** ESTRELA, TAQUARI, VENÂNCIO AIRES **964** BAGÉ, DOM PEDRITO **965** CACHOEIRA DO SUL, CAÇAPAVA **966** RIO PARDO, PÂNTANO GRANDE **968** SANTA CRUZ DO SUL	

E Sintaxe utilizada

```
    # Stepwise
stepwise <- step(mod, direction = "both")


    # Regressão Logística
reg_logistica <- glm(stepwise$formula, family = "binomial", data = dados_analise)


    # Regressão Logística Geograficamente Ponderada
RLGP.spdf <- SpatialPointsDataFrame(dados_analise_RLGP[, 65:66], dados_analise_RLGP)
    DM <- gw.dist(dp.locat = coordinates(RLGP.spdf))


    bw.ggwr1 <- bw.ggwr(reg_logistica$formula, RLGP.spdf, family ="binomial",
approach="AIC", kernel="gaussian", adaptive = FALSE, dMat = DM)


    mod1 <- ggwr.basic(reg_logistica$formula, RLGP.spdf, bw = bw.ggwr1, family
="binomial", kernel = "gaussian", adaptive = FALSE, dMat = DM)


    bw.ggwr2 <- bw.ggwr(reg_logistica$formula, RLGP.spdf, family ="binomial",
approach="AIC", kernel="gaussian", adaptive = TRUE, dMat = DM)


    mod2 <- ggwr.basic(reg_logistica$formula, RLGP.spdf, bw = bw.ggwr2, family
="binomial", kernel = "gaussian", adaptive = TRUE, dMat = DM)


    # Aplicando o modelo na mostra de teste
predito <- predict(reg_logistica, type = 'response', newdata = dados_teste[-(62:64)])


    # Matriz de confusão
previsoes <- ifelse(predito >= 0.6, 1, 0)
matriz_confusao <- table(dados_teste$tp60_atu, previsoes)
```

```
    # AUC
roc <- roc(dados_teste$tp60_atu, predito)
auc <- round(auc(dados_teste$tp60_atu, predito),4)


    # Teste KS
perf <- performance(pred,"tpr","fpr")
ks <- max(attr(perf, "y.values")[[1]] - (attr(perf, "x.values")[[1]]))
```