UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Uma Abordagem Bayesiana em Modelos de Risco de Crédito

Erick Luciano Floriano Mendes

Dissertação de Mestrado do Programa de Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria (MECAI)

ICMC USP SÃO CARLOS

Erick Luciano Floriano Mendes

Uma Abordagem Bayesiana em Modelos de Risco de Crédito

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Mestre – Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria. *VERSÃO REVISADA*

Área de Concentração: Matemática, Estatística e Computação

Orientador: Prof. Dr. Adriano Kamimura Suzuki

USP – São Carlos
Maio de 2022

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

Fa Floriano Mendes, Erick Luciano
Uma Abordagem Bayesiana em Modelos de Risco de Crédito / Erick Luciano Floriano Mendes; orientador Adriano Kamimura Suzuki. -- São Carlos, 2022.
69 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2022.

1. Modelagem de Crédito. 2. Inferência Bayesiana. 3. Modelos Power Link. 4. Risco de Crédito. I. Kamimura Suzuki, Adriano, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Erick Luciano Floriano Mendes

A Bayesian Approach in Credit Risk Models

Master dissertation submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP, in partial fulfillment of the requirements for the degree of the Master – Professional Masters in Mathematics, Statistics and Computing Applied to Industry. *FINAL VERSION*

Concentration Area: Mathematics, Statistics and Computing

Advisor: Prof. Dr. Adriano Kamimura Suzuki

USP – São Carlos
May 2022

Dedico esta pesquisa a todos os pesquisadores brasileiros, que conseguem superar adversidades e, mesmo sem grandes investimentos e incentivos à pesquisa no país, se destacam globalmente.

AGRADECIMENTOS

Agradeço à Universidade de São Paulo e ao Instituto de Ciências Matemáticas e de Computação pela oportunidade e infraestrutura concedida a mim para o desenvolvimento desta pesquisa, bem como todos os colaboradoes destas instituições que me deram suporte nesta jornada.

À Universidade Estadual de Campinas e ao Instituto de Matemática, Estatística e Computação Científica pela minha sólida formação e base do meu conhecimento em Estatística.

Ao meu orientador Adriano Suzuki, por todos os ensinamentos, sugestões e direcionamentos dados ao longo do projeto, além de sua amizade e conselhos durante este período.

Aos meus amigos, que fazem parte da minha história e tornaram essa jornada mais leve.

À minha esposa Giovana, que sempre me apoia e me incentiva em todos os meus objetivos.

À minha família que é e sempre será a minha base, em especial aos meus pais. Minha mãe, que abriu mão de muitas coisas na vida para que eu chegasse até aqui, e meu pai, um exemplo de pessoa que me inspira por seu caráter e sempre me incentivou a ir mais longe.

“Ninguém ignora tudo. Ninguém sabe tudo. Todos nós sabemos alguma coisa. Todos nós ignoramos alguma coisa. Por isso aprendemos sempre.,
”

(Paulo Freire)

RESUMO

MENDES, E. L. F. **Uma Abordagem Bayesiana em Modelos de Risco de Crédito**. 2022. 69 p. Dissertação (Mestrado – Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria) – Institute of Mathematical and Computer Science, University of São Paulo, São Carlos – SP, 2022.

A decisão de aprovação ou não de uma proposta de crédito resulta basicamente em duas opções: ou o crédito é aprovado, ou o crédito é reprovado. Para a aprovação deste crédito as empresas utilizam usualmente o modelo de regressão logística com a estimação dos parâmetros baseada no Estimador de Máxima Verossimilhança, uma técnica considerada da Inferência Clássica, o que limita a usabilidade deste estimador. Uma das abordagens desafiantes muito discutida em pesquisas acadêmicas é a Inferência Bayesiana, em que os parâmetros dos modelos são interpretados como variáveis aleatórias com distribuições definidas a *priori*. Sendo assim, a proposta desta pesquisa foi a utilização de técnicas provenientes da Inferência Bayesiana para avaliar possíveis ganhos que essa abordagem poderia trazer frente à metodologia Clássica. As análises foram desenvolvidas a partir de uma base de dados com cerca de cem mil registros contendo informações da performance de crédito de uma instituição financeira e variáveis preditoras com informações de débitos,consultas, informações geográficas e cadastrais em todo o mercado de crédito. Em posse destas informações, foram testadas abordagens Bayesianas para a estimativa dos parâmetros do modelo, avaliando os resultados em termos de KS e AUC. Avaliou-se também o ganho que as transformações *Power Link* na função e ligação logito poderiam trazer. Foram testados mais de 60 modelos Bayesianos diferentes, com resultados de KS e AUC bastante próximos aos resultados utilizando Inferência Clássica(melhor resultado de KS foi 26.8% e o melhor resultado de AUC foi de 33.0%). Sendo assim, ao final da pesquisa foi possível encontrar modelos Bayesianos com poder discriminante (KS e AUC) próximas ao modelo Clássico, porém com a grande vantagem de obter parâmetros agora com distribuições de probabilidade conhecidas.

Palavras-chave: Modelagem de Crédito, Inferência Bayesiana, Modelos *Power Link* ,Risco de Crédito.

ABSTRACT

MENDES, E. L. F. **A Bayesian Approach in Credit Risk Models**. 2022. 69 p. Dissertação (Mestrado – Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria) – Institute of Mathematical and Computer Science, University of São Paulo, São Carlos – SP, 2022.

The decision of a credit proposal results basically in two options: the credit is approved or the credit is rejected. For this credit approval, companies usually use the logistic regression model with the parameters estimation based on the Maximum Likelihood Estimator, a classical inference technique, that limits the usability of this estimator. One of the challenging approaches much discussed in academic research is Bayesian Inference, which the parameters of the models are interpreted as random variables with distributions previously defined. Therefore, the purpose of this research was use techniques from Bayesian Inference to evaluate possible gains that this approach could bring against the Classical methodology. The analyzes were developed from a database with about one hundred thousand records containing information on the credit performance of a financial institution and predictor variables with information on debts, queries, geographic and registration information throughout the credit market. With this information, Bayesian approaches were tested to estimate the model parameters, evaluating the results in terms of KS and AUC. It was also evaluated the gain that the transformations *Power Link* in the link function logito could bring. More than 60 different Bayesian models were tested, with KS and AUC results very close to the results using Classical Inference (best KS result was 26.8% and the best AUC result was 33.0%). Thus, at the end of the research it was possible to find Bayesian models with discriminating power (KS and AUC) close to the Classic model, but with the great advantage of obtaining parameters now with known probability distributions.

Keywords: Credit Modeling, Bayesian Inference, Power Link Models, Credit Risk.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

1.1 O Mercado de Crédito

O mercado de crédito no Brasil é constituído por instituições financeiras e não-financeiras que prestam serviços de intermediação de recursos para indivíduos e empresas. O acesso ao crédito constitui-se em uma ferramenta fundamental para que indivíduos e empresas possam satisfazer sua capacidade produtiva e, com isso, estimular o crescimento econômico. Sendo assim, o crédito possui hoje um importante papel na economia, uma vez que é essencial ao financiamento do consumo das famílias e do investimento dos setores produtivos, o que possibilita um aperfeiçoamento em aspectos tecnológicos, de estrutura e a geração de empregos, ocasionando a melhoria de vida de diversas pessoas e do país como um todo(DIEESE, 2014).

Segundo o Relatório de Economia Bancária de 2020, gerado anualmente pelo Banco Central do Brasil, atualmente no Brasil ainda há uma concentração do mercado de crédito entre os bancos com maiores volumes de carteira de crédito. As cinco maiores instituições em volume contratado, concentraram em 2020 81,8% do total das operações de crédito registradas, considerando o segmento de bancos comerciais no Banco Central do Brasil. Ao considerar segmentos bancários e não-bancários, esse percentual cai para 68,5%. Pode-se considerar que há moderada concentração do mercado nestas empresas, mas que está caindo: no segmento de bancos comerciais, em 2018 os cinco maiores bancos representavam 84,8% e em 2019 e 83,7%. Considerando segmento bancário e não bancário conjuntamente em 2018 e 2019, esses percentuias foram de 70,9% e 69,8%, respectivamente. Além disso, outro dado relevante sobre o mercado de crédito em 2020 é o volume total de crédito ofertado, acima de R$ 4 trilhões, que representou um crescimento de 15,5% em relação ao ano anterior (BCB, 2020). Tanto a queda na concentração do mercado quanto o crescimento na oferta de crédito não estão relacionado à um único fator, mas alguns fatores que ajudam a entender este fenômeno é o fato de se ter uma maior competitividade entre as empresas do segmento e mais informações disponíveis, que possibilitem um crédito mais seguro.

Com relação à maior competitividade, pode-se destacar o crescimento das empresas chamadas *fintechs*. Não há um conceito único e consensual de *fintech*. Do ponto de vista dos clientes e usuários, as *fintechs* são sinônimo de empresas de pequeno porte que operam por meio de plataformas digitais, fornecendo serviços financeiros diferenciados, de baixo custo, fácil acesso e com apelo tecnológico. Para o regulador financeiro, o conceito é mais amplo e é aonde se encontra um grande desafio, tanto no Brasil quanto no mundo, de como enquadrar essas empresas de maneira a garantir a competitividade no mercado. Regulamentadas pela Resolução 4.656, de 26 de abril de 2018, as Sociedades de Crédito Direto (SCD) e as Sociedades de Empréstimo entre Pessoas (SEP) – as *fintechs* de crédito – iniciaram suas atividades no mercado a partir de 2019. Essa resolução foi o primeiro passo para conseguir regular esse novo formato de empresas. Deste então, a partir desta resolução, a quantidade de empresas SCD e SEP não parou de crescer, chegando a 42 SCDs e 9 SEPs em dezembro de 2020. A quantidade de pedidos de autorização também não para de aumentar: em 31 de dezembro de 2020, encontravam-se em análise 31 pleitos de SCDs e 2 de SEPs, indicando interesse do mercado em atuar nesses segmentos (BCB, 2020).

É também importante entender a função dos *bureaus* de crédito e como eles auxiliam esse mercado de uma maneira geral. Essas empresas coletam informações sobre o histórico de crédito do consumidor no mercado. Desta forma, o consumidor pode se beneficiar com bom comportamento de crédito e as instituições financeiras minimizam o risco, evitando que as pessoas assumam mais dívidas do que podem pagar. A partir da lei do Castro Positivo, de agosto de 2019, informações positivas sobre o comportamento de crédito dos consumidores passaram a ser disponibilizadas para essas empresas de maneira obrigatória, o que beneficia todo o mercado de crédito: com mais informações disponíveis, melhor a assertividade no momento da concessão do crédito, o que pode gerar mais pessoas com acesso a crédito e menores taxas de juros praticadas pelo mercado. Além disso, outras iniciativas estão sendo desenvolvidas e já aplicadas com o mesmo objetivo de compartilhamento de informações entre instituições financeiras com a expectativa de tornar o mercado mais competitivo e com menores riscos de inadimplência. É o caso do Sistema Financeiro Aberto (mais conhecido como *Open Banking*), em que o cliente pode optar por compartilhar seu histórico de pagamentos e informações entre um banco e outro para poder escolher qual a oferta de crédito mais interessante. Todas essas informações tendem a ser utilizadas principalmente como variáveis preditoras de modelos estatísticos que têm o objetivo de predizer a probabilidade de inadimplência.

No desenvolvimento de modelos estatísticos para concessão de crédito é bastante usual a utilização de modelos de regressão logística, considerando estimação dos parâmetros via inferência clássica, com estimadores de máxima verossimilhança. Além disso, nos últimos anos houve um crescimento na aplicação de modelos de Aprendizado de Máquina para discriminar bons e maus pagadores, mas que não serão avaliados nesta pesquisa. Para a construção dos modelos utiliza-se a performance observada para os consumidores que tiveram propostas de crédito aprovadas, ou seja, considera-se o comportamento de atraso do público apenas na empresa

que concedeu o crédito. Entende-se que este é o dado mais seguro do comportamento de crédito especificamente para uma determinada empresa, portanto são os dados mais recomendáveis para esse público.

Se por um lado, modelos de Aprendizado de Máquina podem trazer benefícios do ponto de vista de poder discriminante (KS), por outro lado, oferece modelos com menor interpretabilidade quando comparado ao modelo de regressão logística, o que pode gerar indisposições com clientes que solicitem entender exatamente o por-quê da nota de crédito que foi atribuída. Então modelos de regressão logística apresentam, em geral, bons patamâres de discriminação, geralmente inferiores aos de Aprendizado de Máquina, porém com uma maior simplicidade na interpretação da probabilidade e dos pesos das variáveis.

O objetivo deste projeto de pesquisa é apresentar alternativas interessantes para desenvolvimento de modelos de crédito, utilizando ferramentas de inferência Bayesiana e avaliar os possíveis ganhos com a utilização destas ferramentas.

1.2 Estudos relacionados

A pesquisa proposta está dentro de uma grande área que é a área de risco de crédito. Ao pesquisarmos os principais artigos relacionados com risco de crédito, podemos destacar o artigo *Bankruptcy prediction for credit risk using neural networks: A survey and new results*, em que o autor propõe a utilização de sistemas com redes neurais para predizer se uma empresa irá falir (ATIYA, 2001). O artigo *A new fuzzy support vector machine to evaluate credit risk* propõe a utilização de máquina de vetores-suporte (SVM) com uma nova abordagem considerando a teoria de Fuzzy para discriminar bons e maus pagadores (WANG; LAI, 2005). As duas pesquisas são as mais citadas relacionadas aos temas de Risco de Crédito, segundo a plataforma IEEE. As pesquisas são bastante relevantes, ainda mais se considerarmos a época da pesquisa, antes de alguns avanços tecnológicos, como a limitação da capacidade computacional e a complexidade dos métodos utilizados.

No artigo *A novel credit scoring model based on optimized random forest*, o autor propõe diversas técnicas de aprendizado de máquina, como SVM, Floresta Aleatória, entre outros, sempre comparando os resultados de cada técnica (ZHANG; YANG; ZHOU, 2018). O artigo *A Novel Reject Inference Model Using Outlier Detection and Gradient Boosting Technique in Peer-to-Peer Lending*, avalia a utilização de *Gradient Boosting* para inferência de rejeitados específicamente para operações de crédito *peer-to-peer*, ou seja, de pessoa para pessoa (XIA, 2019).

No artigo *Credit Scoring Model based on Kernel Density Estimation and Support Vector Machine for Group Feature Selection*, de 2018, os autores propõem novamente uma abordagem considerando SVM para melhorar a performane dos modelos de crédito (ZHANG; ZHOU, 2018). No artigo *Bound and collapse Bayesian reject inference for credit scoring*, os autores propõem

um método flexível para gerar as probabilidades a partir de um modelo baseado na técnica de colapso e fronteira Bayesianos, e obtêm resultados positivos, com maior poder de classificação que os métodos usuais (CHEN; ÅSTEBRO, 2011).

No artigo (LEONG, 2016), o autor propõe um modelo de redes Bayesianas para endereçamento de censura, com classes desbalanceadas e ainda no contexto de risco de crédito. Ao comparar os resultados com outros modelos, como Redes Neurais Artificiais e Regressão Logística, o autor conclui que o modelo proposto apresenta melhor desempenho com relação as métricas de acurácia, sensibilidade, precisão e curva ROC. Outro artigo relacionado ao tema é (WILHELMSEN M., 2009) em que o autor aplica o método de Aproximação de Laplace Aninhada Integrada (da sigla em inglês, INLA) para modelagem Bayesiana de risco de crédito.No artigo, o autor demonstra como o método INLA fornece uma rápida e precisa inferêcia para modelos de risco de crédito e compara sua performance com a estimação Bayesiana considerando Monte Carlo via Cadeia de Markov.

1.3 Organização da Dissertação

No Capítulo 2 desta dissertação, discute-se as técnicas e métricas de avaliação utilizadas na pesquisa proposta, bem como uma introdução sobre a base de dados utilizada. Na sequência, no Capítulo 3, apresenta-se as aplicações e os resultados encontrados a partir da base de dados do estudo. Por fim, no Capítulo 4 apresenta-se uma discussão sobre os resultados das metodologias aplicadas e as conclusões a respeito da abordagem proposta.

CAPÍTULO

2

MATERIAIS E MÉTODOS

Este Capítulo tem como objetivo apresentar as principais abordagens, teorias, métodos e métricas de avaliação utilzadas para o desenvolvimento da pesquisa, bem como uma explicação sobre o banco de dados utilizado. Na subseção 2.1 são apresentados os principais conceitos com realação à Inferência, sendo ela Clássica ou Bayesiana, além dos estimadores que foram utilizados. Na subseção 2.2 apresenta-se o modelo de Regressão Logística e a transformação *Power Link*. Na subseção 2.3 as métricas de avaliação Teste de Geweke, Kolmogorov-Smirnov e Área sob a Curva ROC são apresentadas. Na subseção 2.4 apresenta-se a base de dados utilizada.

2.1 Inferência

Segundo Schervish e DeGroot (2014) Inferência Estatística é o processo que produz uma afirmação probabilística sobre alguma ou todas as partes de um modelo estatístico, podendo interpretar uma "afirmação probabilística"como uma afirmação que faz uso de qualquer conceito da teoria de probabilidade: média, média condicional, quantis e variância são alguns exemplos. Existem diversas abordagens que podem ser utilizadas para a realização destes processos estatísticos definidos. Porém, antes de uma breve explicação sobre as abordagens utilizadas nesta pesquisa, é importante esclarecer algumas definições, que são independentes da abordagem utilizada. Essas definições podem ser encontradas em Schervish e DeGroot (2014).

Parâmetro: uma característica ou combinação de características que determinam a distribuição conjunta de uma variável aleatória de interesse, que será chamada de parâmetro da distribuição.

Distribuição *a Priori*: distribuição de um parâmetro antes da observação de qualquer dado. Suponha um modelo estatístico com o parâmetro θ. Se este modelo tratar θ como aleatório, então esta será a distribuição que este modelo atribui a θ antes de observar as demais variáveis aleatórias de interesse, chamada de distribuição *a priori* de θ, denotada por $f(\theta)$. Vale destacar

que pode-se generalizar θ também para um vetor de parâmetros $\underset{\sim}{\theta}$, e assim, obter $f(\underset{\sim}{\theta})$ sendo a distribuição *a priori* de $\underset{\sim}{\theta}$.

Distribuição a *Posteriori*: A distribuição de um parâmetro condicionado aos dados observados é chamada de distribuição *a posteriori*. Suponha um problema de inferência estatística com o parâmetro θ e as variáveis aleatórias $X_1, X_2, ..., X_n$ que serão observadas. A distribuição condicional de θ dado $X_1, X_2, ..., X_n$ é chamada de distribuição a posteriori de θ, denotada por $f(\theta|\underset{\sim}{X})$, onde $\underset{\sim}{X} = X_1, X_2, ..., X_n$.

Função de Verossimilhança: Ao se conectar os valores observados do conjunto de dados como as funções densidade de probabilidade condicionais ou funções massa de probabilidade condicionais (dos dados condicionado ao parâmetro θ), o resultado é uma função do parâmetro sozinho, chamada de função de verossimilhança, denotada por $f(\underset{\sim}{X}|\theta)$. Considere $X_1, X_2, ..., X_n$ variáveis aleatórias independentes com funções densidade ou massa de probabilidade dadas por $f(x_1|\theta), f(x_2|\theta), ..., f(x_n|\theta)$, em que θ é o parâmetro desconhecido. Sendo assim, a Função de verossimilhança será dada por:

$$L(x|\theta) = f(x_1|\theta)f(x_2|\theta)...f(x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta).$$

A função de verossimilhança é a função densidade ou massa de probabilidade de probabilidade condicional dos dados dado um parâmetro. A verossimilhança mostra o quanto os dados irão alterar uma incerteza. Valores altos de verossimilhança correspondem a valores de parâmetros em que a função densidade de probabilidade da *posteriori* será maior que a da *priori*. Valores baixos de verossimilhança ocorrem em valores de parâmetros onde a *posteriori* será menor que a *priori*.

Existem diversas abordagens classificadas dentro dos processos de Inferência Estatística. Entre elas, temos a Inferência Clássica e a Inferência Baysiana, que foram aplicadas nesta pesquisa.

A abordagem Clássica, também conhecida como Frequentista, considera que o parâmetro desconhecido θ é fixo, não possuindo uma função de probabilidade atribuída, e é baseada no princípio de amostragem repetida, que supõe ser possível realizar infinitas vezes um experimento, portanto o x observado é apenas um dos possíveis resultados que o experimento poderia ter. O Estimador de Máxima Verossimilhança, que será explicado a seguir, é uma das técnicas mais utilizadas dentro da abordagem Clássica.

Na abordagem Bayesiana, o parâmetro desconhecido θ assume uma distribuição de probabilidade. Desta forma, θ é tratado como uma variável aleatória com uma distribuição a priori. Com isso o estimador de Bayes pode ser dado a partir de alguns conceitos da teoria da probabilidade, como a média, a mediana, a moda ou algum quantil da distribuição *a posteriori* do parâmetro.

2.1.1 Estimador de Máxima Verossimilhança

A estimação via Máxima Verossimilhança é um método de estimação e de escolha do estimador dos parâmetros, considerado como um método da Inferência Clássica, que evita a utilização de distribuições *a priori*. Ele escolhe como a estimativa de θ o valor que maximiza a função de verossimilhança. Para cada possível observação do vetor x, seja $\delta(x) \in \Omega$ denotando um valor de $\theta \in \Omega$, no qual a função de verossimilhança $f(x|\theta)$ assume o valor máximo. Seja $\hat{\theta} = \delta(X)$ o estimador de θ definido desta maneira. Este estimador de θ é chamado de Estimador de Máxima Verossimilhança de θ. Após ser observado $X = x$, o valor de $\delta(x)$ é chamado de Estimativa de Máxima Verossimilhança de θ. O processo para encontrar este ponto de máximo é dado por:

1. Encontrar a Função de Verossimilhança ($L(\underset{\sim}{X}|\theta)$);

2. Calcular a primeira derivada de $L(\underset{\sim}{X}|\theta)$ em relação a θ. Encontrar os possíveis valores de θ que zeram esta função;

3. Calcular a segunda derivada de $L(\underset{\sim}{X}|\theta)$ e aplicar os valores de θ encontrados na etapa anterior. Aquele que corresponder a um valor negativo para a função da segunda derivada de $L(\underset{\sim}{X}|\theta)$ é o ponto que maximiza a verossimilhança, ou seja, o Estimador de Máxima Verossimilhança.

Agora serão apresentadas algumas propriedades para este tipo de estimador, que foram importantes para a sua ampla utilização.

Invariância: Se $\hat{\theta}$ é o Estimador de Máxima Verossimilhança de θ e g é uma função injetiva, então $g(\hat{\theta})$ é o Estimador de Máxima Verossimilhança de $g(\theta)$.

Consistência: Considere um problema de estimação em que uma amostra aleatória é retirada de uma distribuição envolvendo um parâmetro θ. Suponha que para cada tamanho amostral n suficientemente grande, ou seja, para cada valor de n maior que algum valor mínimo, existe um único Estimador de Máxima Verossimilhança de θ. Então, sob certas condições, as quais são tipicamente satisfeitas em problemas práticos, a sequência de Estimadores de Máxima Verossimilhança é uma sequência consistente de estimadores de θ. Em outras palavras, em alguns problemas, a sequência de Estimadores de Máxima Verossimilhança converge em probabilidade para o valor desconhecido de θ a medida que n converge para o infinito. Isso indica também que esses estimadores são não-viciados para tamanhos de amostra suficientemente grandes.

Eficiência Assintótica: Segundo o Teorema do Limite Inferior de Cramer-Rao, para um dado parâmetro qualquer, existe um limite inferior para a variância das estimativas não-viciadas. Para grandes amostras, os Estimadores de Máxima Verossimilhança atingem esse limite e, portanto, têm a menor variância possível dentre as estimativas não-viciadas.

2.1.2 Inferência Bayesiana

Inferência Bayesiana é um processo de ajuste de um modelo de probabilidade para um conjunto de dados, que resume os resultados por uma distribuição de probabilidade para os parâmetros do modelo e quantidades não-observáveis como predições para novas observações.

A característica essencial dos métodos Bayesianos é o uso explícito de probabilidades para quantificar incertezas em inferências baseadas em análises estatísticas.

Em Gelman *et al.* (1995), o processo de análise Bayesiana é dividido em três passos:

1. Configurar um modelo completo de probabilidade - uma distribuição de probabilidade conjunta para todas as quantidades observáveis e não-observáveis existentes no problema;

2. Condições sobre os dados observáveis: calcular e interpretar a distribuição *a posteriori* apropriada - a distribuição de probabilidade condicional sobre as quantidades não-observáveis de interesse, a partir de dados observáveis;

3. Avaliação do ajuste do modelo e as implicações dos resultados da distribuição *a posteriori*: avaliar se o modelo se ajusta bem aos dados.

Uma motivação inicial para o pensamento Bayesiano é a sua facilidade em interpretar um senso comum em conclusões estatísticas. Por exemplo, um intervalo Bayesiano para uma quantidade de interesse desconhecida pode ser diretamente considerado como tendo uma alta probabilidade de conter a quantidade desconhecida, em contraste com a inferência clássica, em que o intervalo de confiança deve ser estritamente interpretado apenas em relação com uma sequência de inferências similares que devem ser feitas em práticas repetidas.

Conclusões Bayesianas acerca do parâmetro de interesse θ, ou dados não-observáveis $\widetilde{y}$, são realizadas em termos de afirmações de probabilidade. Essas afirmações são condicionadas aos valores observados y, podendo utilizar as notações $p(\theta|y)$ e $p(\widetilde{y}|y)$.

2.1.2.1 Regra de Bayes

Com o objetivo de realizar afirmações de probabilidade sobre θ dado y, deve-se inicar um modelo fornecendo a probabilidade conjunta de θ e y. A função densidade ou massa da probabilidade conjunta pode ser escrita como o produto de duas distribuições que geralmente se referem à distribuição *a priori* (distribuição $p(\theta)$) e a distribuição amostral dos dados ($p(y|\theta)$, também chamada de verossimilhança) respectivamente:

$$p(\theta|y) = \frac{p(\theta,y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)},$$

em que $p(y) = \sum_\theta p(\theta)p(y|\theta)$, sendo a soma considerando todos os possíveis valores de θ. Omitindo o fator $p(y)$, que não depende de θ e que pode ser considerado uma constante se y for fixo, pode-se produzir a função de densidade *a posteriori* não-normalizada descrita abaixo:

$$p(\theta|y) \propto p(\theta)p(y|\theta)$$

2.1.2.2 Predição Bayesiana

Ao realizar inferências acerca de uma observação deseconhecida, segue-se a mesma lógica. Antes de considerar os dados de y, a distribuição de y desconhecido mas observável é, para o caso em que θ é contínua :

$$p(y) = \int p(y,\theta)d\theta = \int p(\theta)p(y|\theta)d\theta,$$

e para o caso em que θ é uma variável aleatória discreta é:

$$p(y) = \sum_\theta p(y,\theta) = \sum_\theta p(\theta)p(y|\theta).$$

Essa distribuição é conhecida como distribuição marginal de y, mas também pode ser chamada de distribuição preditiva *a priori*: *priori* pois não é condicionada a nenhuma observação prévia do processo e preditiva pois é a distribuição de uma quantidade que é observável.

Para definição das distribuições *a priori*, pode-se considerar dois casos: prioris informativas e não-informativas. Prioris informativas são utilizadas quando o pesquisador possui algum conhecimento prévio sobre os parâmetros e deseja utilizar esse conhecimento para a realização da inferência. No caso das prioris não-informativas, ou não há, ou há pouca informação prévia disponível, ou ainda há, mas não deseja-se utilizar essa informação para a realização da inferência.

Sendo assim, pode-se dizer que a inferência Bayesiana necessita da distribuição amostral (proveniente da verossimilhança) e uma distribuição *a priori* para o parâmetro. De posse destas distribuições, encontra-se a distribuição *a posteriori* dos parâmetros, objeto de todos os procedimentos com a abordagem Bayesiana acerca dos parâmetros de interesse. No estudo realizado para esta pesquisa, optou-se por utilizar o estimador Bayesiano como a média da distribuição *a posteriori* dos parâmetros.

Para entendimento mais detalhado sobre Inferência Bayesiana, pode-se consultar Gelman *et al.* (1995).

2.2 Regressão Logística

O Modelo de Regressão Logística é um Modelo Linear Generalizado que fornece um método adequado para modelagem de uma variável resposta binária, em que essa variável resposta

Y_i segue a distribuição de probabilidade Bernoulli, assumindo o valor "1" com probabilidade π_i e "0" com probabilidade $1-\pi_i$. Sendo assim, i varia de acordo com as observações, ou seja, $i = 1,2,...,n$, sendo resultante da função inversa da logística sobre um vetor $\underset{\sim}{x_i}$, que inclui uma constante e outras $k-1$ variáveis explicativas (KING G., 2001).

A distribuição de Bernoulli possui função de probabilidade da seguinte forma:

$$\mathbb{P}(Y_i|\pi_i) = \pi_i^{Y_i}(1-\pi_i)^{(1-Y_i)}.$$

Os parâmetros desconhecidos estão presentes no vetor $\underset{\sim}{\beta}$, um vetor $k \times 1$ com o primeiro elemento representando uma constante e os demais sendo os parâmetros correspondentes a cada variável explicativa do modelo. Esses são os parâmetros de interesse em que deseja-se gerar estimadores e estimativas. Nesta pesquisa, optou-se por utilizar o Estimador de Máxima Verossimilhança como o estimador que representará a Inferência Clássica, e o Estimador de Bayes proveniente da média da distribuição *a posteriori* dos parâmetros, representando a Inferência Bayesiana.

Uma maneira alternativa de definir o mesmo modelo é supondo uma variável contínua não observável Y_i^* distribuída de acordo com a função de densidade da logística com média μ_i. Desta maneira, μ_i varia de acordo com o vetor de observações como uma função linear de $\underset{\sim}{x_i}$. O modelo seria bastante próximo de uma regressão linear se Y_i^* fosse observável.

Sendo assim, teríamos $Y_i^* \sim Logistica(Y_i^*|i)$ e com a seguinte função densidade de probabilidade:

$$\mathbb{P}(Y_i^*) = \frac{e^{-(Y_i^*-\mu_i)}}{(1+e^{-(Y_i^*-\mu_i)})^2},$$

podendo ser escrito da seguinte forma:

$$\mathbb{P}(Y_i^* = 1|\underset{\sim}{\beta}) = \pi_i = \mathbb{P}(Y_i^* > 0|\underset{\sim}{\beta}) = \frac{1}{1-e^{\underset{\sim}{x_i}\underset{\sim}{\beta}}}.$$

A regressão logística é uma abordagem simples para resolução de problemas de classificação, porém, em geral apresenta resultados bastante satisfatórios, com uma grande vantagem da interpretabilidade do modelo. Além disso, é amplamente utilizada no mercado de crédito. No estudo em questão, a regressão logística foi aplicada para a predição de clientes bons pagadores.

2.2.1 Power Links

O modelo de Regressão Logística apresentado na subseção anterior pode ser reescrito da seguinte forma:

$$\mathbb{P}(Y_i^* = 1|\beta) = L(\underset{\sim}{X}|\theta) = \frac{1}{1-e^{\underset{\sim}{x_i}\underset{\sim}{\beta}}},$$

em que $L(\underset{\sim}{X}|\theta)$ representa a função de distribuição acumulada de uma distribuição Logística padrão. A função inversa $L^{-1}(\underset{\sim}{X}|\theta)$ é chamada função de ligação na qual $\underset{\sim}{x_i}\,\underset{\sim}{\beta}$ é o preditor linear correspondente.

Considere F^{-1} funções de ligação para regressão binária com $F(.)$ e $f(.)$ sendo as funções de distribuição acumulada (fda) e densidade de probabilidade (fdp), respectivamente, de uma variável aleatória S. Podendo assim definir a propriedade reversa e as distribuições de potência (*power*) e potência reversa (*reversal power*)(BAZáN *et al.*, 2017).

Definição 1

Seja $S \sim F(.)$. Pode-se dizer que a distribuição de S satistaz a propriedade reversa se fda de -S é uma distribuição diferente que pode ser escrita como -$S \sim G(.) \equiv 1$ - $F(-.)$. Neste caso, a função $G(.)$ é chamada de distribuição reversa de $F(.)$.

Resultado 1

Considere $S \sim F(.)$ uma variável aleatória que segue uma distribuição simétrica, ou seja, uma fdp $f(.)$ na qual existe um valor x_0 tal que $f(x_0 - \delta) = f(x_0 + \delta)$ para todos os números reais δ. Então, $F(.)$ não satisfaz a propriedade reversa pois S e $-S$ possuem as mesmas fda e fdp. As distribuições Logística, Normal e Cauchy são alguns exemplos (BAZáN *et al.*, 2017).

Definição 2

Pode-se dizer que uma variável aleatória univariada S segue uma distribuição de potência $S \sim P(\mu, \sigma^2, \lambda)$ com parâmetros de localização, escala e forma dados por -$\infty < \mu < \infty$, $\sigma^2 > 0$ e $\lambda > 0$, respectivamente, se a densidade desta distribuição possui a seguinte forma:

$$f_p(s|\mu, \sigma^2, \lambda) = \frac{\lambda}{\sigma} g\left(\frac{s-\mu}{\sigma}\right)\left[G\left(\frac{s-\mu}{\sigma}\right)\right]^{\lambda-1},$$

com $G(.)$ denotando uma fda absoluta qualquer com suporte nos reais e $g(.)$ é uma fdp unimodal com log-concavidade e com suporte na linha dos reais (-∞,∞) chamada de distribuição de linha de base.

Se $\lambda = 1$, a densidade de S na equação acima apresentada se reduz à densidade de $g(\mu, \sigma^2)$. A distribuição de potência padrão é obtida quando $\mu = 0$ e $\sigma^2 = 1$, sendo denotada por $Z \sim P(\lambda)$, com as seguintes fda e fdp, respectivamente:

$$f_p(z|\lambda) = \lambda g(z)[G(z)]^{\lambda-1}$$

$$F_p(z|\lambda) = [G(z)]^{\lambda}$$

Diversas transformações de potência na função de ligação, mais conhecidas pelo nome em inglês *Power Links*, podem ser aplicadas. No contexto desta pesquisa é proposta a transformação da função de ligação da regressão logística, a função logito. Para isso foram testadas a *Power*

Link padrão e a *Power Link* Reversa, apresentadas na Tabela 1, que transformam a distribuição da função logito em uma distribuição assimétrica.

Tabela 1 – *Power Links* avaliadas

Nome da função de ligação	função de ligação	*Power Link*	*Power Link* reversa
logito	$L(z) = \frac{1}{(1+e^{-z})}$	$L(z)^{\lambda}$	$(1-L(-z))^{\lambda}$

2.3 Métricas de Avaliação

2.3.1 Teste de Geweke

Os métodos Monte Carlo via Cadeia de Markov (MCMC) permitem obter uma amostra da distribuição *a posteriori* de interesse, que não poderia ser simulada diretamente. Para isso são gerados valores de forma iterativa a partir de cadeias de Markov. A ideia geral destes métodos é simular um passeio aleatório no espaço do parâmetro de interesse θ, que converge para uma distribuição estacionária, que é a distribuição *a posteriori* $f(\theta|y)$, em que $\underset{\sim}{y}$ é o vetor de observações.

No processo MCMC, a variável gerada numa etapa depende da variável gerada na etapa anterior, de modo que a série tende a apresentar uma autocorrelação. Para se formar uma amostra que resulte no parâmetro de interesse é necessário que o processo tenha alcançado estacionariedade, o que pode não acontecer se houver um limite na quantidade de iterações a serem testadas.

Para avaliar se a estacionaridade foi alcançada pode ser utilizado o teste de Geweke. Esse teste baseia-se em testar a igualdade das médias da primeira e da última parte da cadeia de Markov. Se as amostras são retiradas da distribuição estacionária da cadeia, as duas médias tendem a ser iguais; sob a hipótese nula a estatística do teste, que indica convergência do método, possui uma distribuição assintótica normal (BORGES, 2008).

A estatística do teste é dada por

$$Z = \frac{\bar{\theta}_1 - \bar{\theta}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}},$$

na qual, $\bar{\theta}_1$ e S_1^2 representam, respectivamente, a média e o estimador da variância das primeiras n_1 observações da cadeia, enquanto $\bar{\theta}_2$ e S_2^2 representam a média e o estimador da variância nas últimas n_2 observações. É possível ver também que a estatística do teste Z converge para uma distribuição $Normal(0,1)$ quando a quantidade de iterações $N \to \infty$.

Para os resultados desta pesquisa, considerou-se as 10% primeiras observações e as 50% últimas observações para a realização deste teste. Além disso, o nível de significânica utilizado foi $\alpha = 0.05$.

2.3.2 Kolmogorov-Smirnov (KS)

Uma medida bastante utilizada no mercado de trabalho para mensurar a qualidade de um modelo de classificação é a Estatística de Kolmogorov-Smirnov, ou simplesmente, KS.

Essa métrica é calculada encontrando o supremo do módulo da diferença entre duas distribuições acumuladas de probabilidade, podendo ser descrita como (KOLáCEK J., 2010):

$$KS = max_a |F_0(a) - F_1(a)|,$$

em que F_0 e F_1 representam as funções de distribuição de probabilidade acumulada de X_0 e X_1, respectivamente. Sendo assim, $0 \leq F_0(a) \leq 1$ e $0 \leq F_1(a) \leq 1$, fazendo com que $0 \leq KS \leq 1$. Quanto mais próximo de 1, maior o poder discriminante do modelo.

2.3.3 Área sob a Curva ROC (AUC-ROC)

Outra métrica bastante utilizada para avaliação de modelos de classificação é a Área sob a Curva ROC. O primeiro passo para entender como essa métrica é calculada é entender o que significa a curva ROC.

a curva ROC (do inglês *Receiver Operating Characteristic* ou Característica de Operação do Receptor) indica o quanto o modelo de classificação avaliado pode discriminar duas classes.

Considere um modelo de classificação binária que prediz qual a classe para cada observação, podendo assumir valores '0' (negativo) ou '1' (positivo). Com isso, podemos definir quatro medidas:

- os Verdadeiros Positivos: observações positivas que foram classificadas pelo modelo como positivas;
- os Falsos Positivos: observações negativas que foram classificadas pelo modelo como positivas;
- os Verdadeiros Negativos: observações negativas que foram classificadas pelo modelo como negativas;
- os Falsos Negativos: observações positivas que foram classificadas pelo modelo como negativas.

A partir dessas medidas pode-se definir a Taxa de Verdadeiro Positivo (TVP) e a Taxa de Falso Positivo (TFP). Essas taxas são diretamente relacionadas às métricas de Sensitividade e Especificidade:

$$TVP = \frac{VerdadeirosPositivos}{VerdadeirosPositivos + FalsosNegativos} = Sensitividade$$

$$TFP = \frac{FalsosPositivos}{FalsosPositivos + VerdadeirosNegativos} = 1 - Especificidade$$

Essas taxas serão calculadas para diversos limiares de classificação, gerando a curva ROC, explicitada na Figura 1.

Com isso é possível calcular o valor da área sob essa curva, resultando no AUC. Na Figura 1 essa área é representada por toda a área preenchida abaixo da curva ROC.

Assim como o KS, o AUC esta entre o intervalo 0 e 1, $0 \leq AUC \leq 1$, sendo que quanto mais próximo de 1 maior o poder de discriminação do modelo. Além disso, vale destacar que um modelo que apresente AUC igual à 0.5 é um modelo que tem o mesmo poder de discriminação de um preditor randômico.

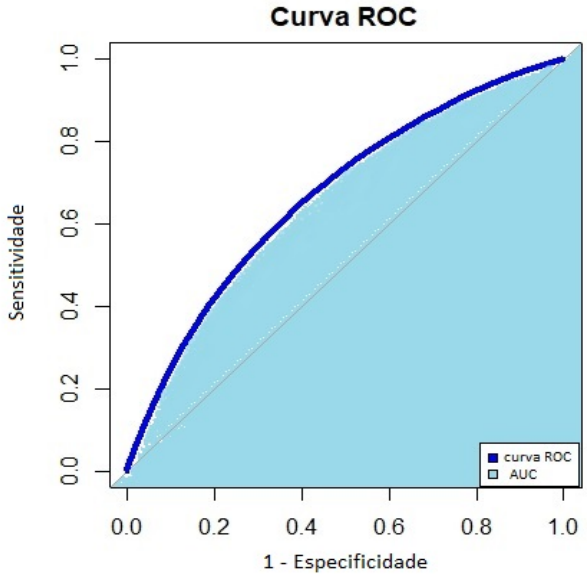

Figura 1 – Curva ROC

Fonte: Elaborada pelo autor.

2.4 Base de Dados

A base de dados utilizada para o desenvolvimento da pesquisa é composta por 97 variáveis e 99.766 registros, sendo 94 variáveis preditoras, 2 variáveis de inferência de performance e uma

variável chave de identificação. Cada registro representa uma proposta aprovada de uma pessoa física para um produto de crédito de um banco digital brasileiro. Essas propostas referem-se ao período entre Junho de 2019 e Dezembro de 2019. Todos os registros da base de dados tiveram suas propostas de crédito aprovadas, o que significa que o banco digital consegue mensurar se o cliente foi um bom pagador com as informações internas do banco (variável "Ind_Bom"). Existe também a marcação de mercado, proveniente de um *bureau* de crédito ("Ind_Bom_merc"), disponível para todos os registros. Para a marcação de performance de crédito foi considerado o conceito *ever* de 60 dias em 6 meses. Nesse conceito, é verificado se na janela de 6 meses posterior à data de concessão do crédito o cliente possui um acúmulo de 60 dias de atraso. Caso possua, o cliente é considerado um mau pagador por esse critério. A variável "Ind_Bom"faz essa avaliação considerando apenas os atrasos internos dos clientes, enquanto a variável "Ind_bom_merc"avalia esse comportamento considerando todos os parceiros de um *bureau* de crédito.

As variáveis preditoras são variavéis provenientes também do *bureau* de crédito, trazendo informações cadastrais, geográficas, sobre o histórico de débitos e de consultas ao mercado de crédito para cada cliente.Todas as variáveis possuem preenchimento, ou seja, não foi necessário a realização de nenhum tratamento para dados faltantes. A Tabela 2 apresenta a quantidade de variáveis preditoras por cada tipo identificado na base de dados. Não é possível saber exatamente o significado de cada variável por suas descrições não estarem disponíveis; apenas é possível identificar qual a classificação de cada uma. Ao se tratar de variáveis do tipo Débito, as informações estarão relacionadas a informações de negativação do cliente do mercado de crédito, podendo variar o tipo de produto, o prazo de observação entre outras coisas. No caso de variáveis de consulta, avalia-se o quanto o consumidor está procurando crédito no mercado, tendo informações de quantas consultas foram realizada por empresas de crédito para aquele consumidor. Variáveis cadastrais são variáveis relacionadas ao cadastro do cliente, como ano de nascimento, estado civil, entre outras. Variáveis geográficas são variáveis relacionadas à região do consumidor, podendo ser o Indice de Desenvolvimento Humano daquele CEP, daquele município, do estado, ou ainda outros indicadores regionais.

Todas as análises, tratamentos e resultados foram gerados a partir do software *R*. Ao final desta dissertação, na seção de Apêndices, estão expostos alguns dos códigos utilizados.

Tabela 2 – Tipos de Variáveis Preditoras

Tipo	**Quantidade**	**Quantidade (%)**
Cadastral	3	3,2%
Consulta	8	8,5%
Débito	39	41,5%
Geográfica	44	46,8%
TOTAL	**94**	**100,0%**

CAPÍTULO

3

APLICAÇÕES

3.1 Pré-Modelagem

Como boas práticas de modelagem, visando a generalização dos modelos adotados, a partir da base de dados do estudo foram geradas duas amostras a partir da técnica de Amostragem Aleatória Simples: uma com 70% da volumetria da base (69.930 registros), chamada de Base de Desenvolvimento, e outra com 30% da volumetria da base (29.836 registros), chamada de Base de Validação. Os parâmetros dos modelos logísticos serão gerados a partir da Base de Desenvolvimento, e os resultados finais serão analisados na Base de Validação, a fim de retirar qualquer tipo de sobreajuste do modelo. A Tabela 3 apresenta a volumetria e a inadimplência das duas amostras, que aparentam ter um comportamento parecido, o que é razoável, dado que tratam-se de amostras aleatórias geradas a partir de uma mesma base de dados.

Tabela 3 – Separação das Amostras de Desenvolvimento e Validação

Amostra	**Volume**	**Volume (%)**	**Inadimplência (%)**
Desenvolvimento	69.930	70,10%	46,8%
Validação	29.836	29,90%	46,6%
Total	**99.766**	**100,00%**	**46,7%**

A princípio, as variáveis preditoras disponíveis são variáveis contínuas, por definição. Sendo assim, para manter essa premissa valida, foram filtradas variáveis que possuem alta concentração ($\geq 90\%$) de volume em um único valor. Das 94 variáveis preditoras, 12 foram filtradas nesta etapa, restando ainda 82 variáveis para o estudo.

A Tabela 4 apresenta as variáveis filtradas nesta etapa. Os nomes das variáveis foram anonimizados com o intuito de preservar as fontes, porém podemos verificar que a maior parte destas variáveis é do tipo Débito.

Tabela 4 – Variáveis com concentração de volume $\geq 90\%$

Variável	**Tipo de Variável**	**Concentração de Volume (%)**
VPF1_0001	Débito	95%
VPF1_0002	Débito	94%
VPF1_0003	Débito	94%
VPF1_0004	Débito	93%
VPF1_0005	Débito	93%
VPF1_0006	Débito	93%
VPF1_0007	Débito	93%
VPF1_0008	Débito	93%
VPF1_0009	Geográfica	93%
VPF1_0010	Consulta	92%
VPF1_0011	Débito	92%
VPF1_0012	Débito	90%

Como o estudo tem o objetivo de avaliar o método de Regressão Logística, e uma das premissas desta técnica é a independência das variáveis preditoras, sendo assim, visando contornar o problema da multicolinearidade, foram filtradas também variáveis que possuíam correlação de Pearson $\geq |0.7|$. Quando duas variáveis possuiam correlação de Pearson $\geq |0.7|$, retirava-se a que possuia maior concentração de volume em um único valor. No total, 18 variáveis possuiam alta correlação com pelo menos mais uma variável. A Figura 2 apresenta as correlações destas 18 variáveis. Nesta etapa, 13 variáveis foram filtradas, enquanto as outras 5 variáveis correlacionadas com as demais foram mantidas para o estudo. A Figura 3 apresenta a correlação final destas 5 restantes.

Após estes filtros, finaliza-se a etapa de "Pré-Modelagem", onde foram aplicados filtros que retiram alguns possíveis vieses. Agora, com a definição das bases de Desenvolvimento e Validação e com as 69 variáveis que não possuem concentração de volume $\geq 90\%$ e não possuem correlação de Pearson $\geq |0.7|$ é possível iniciar a etapa de Modelagem, onde serão confrontadas abordagens Clássica e Bayesianas.

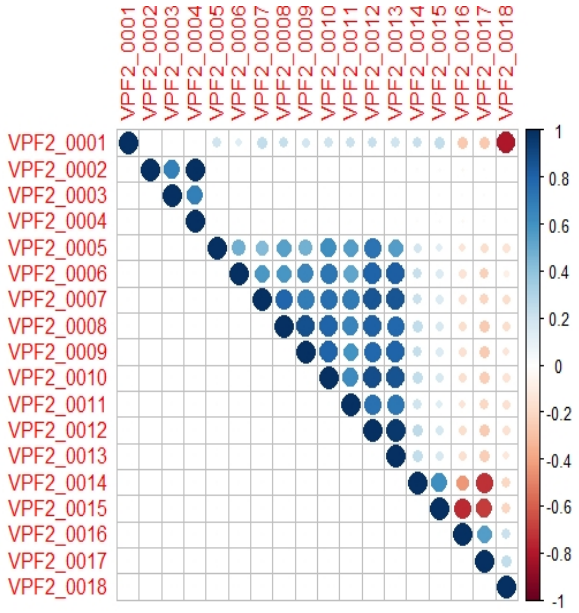

Figura 2 – Variáveis com Correlação $\geq |0.7|$

Fonte: Elaborada pelo autor.

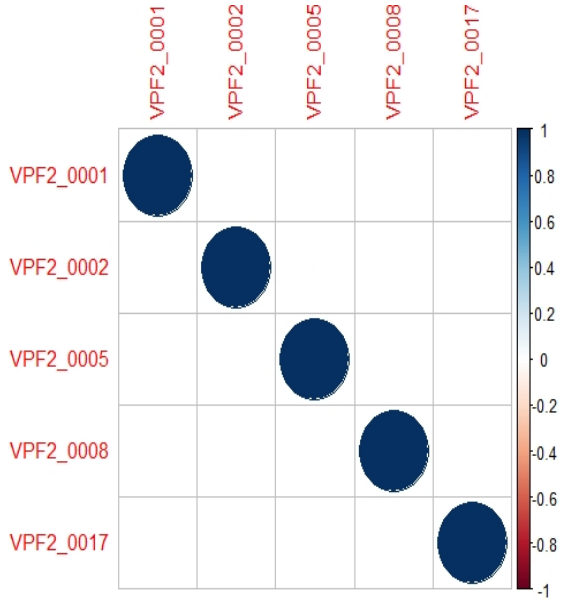

Figura 3 – Variáveis com correlação mantidas no estudo

Fonte: Elaborada pelo autor.

3.2 Modelagem

Nesta subseção serão apresentados os resultados da metodologias desenvolvidas e ilustrados pela Figura 4. Inicialmente foi realizado o Pré-Processamento da base de dados, seguindo

para a realização da Modelagem via estimação Clássica, dando continuidade com a Regressão Logística Bayesiana com prioris Não-informativas e também com transformações *Power Link* finalizando com as conclusões do estudo.

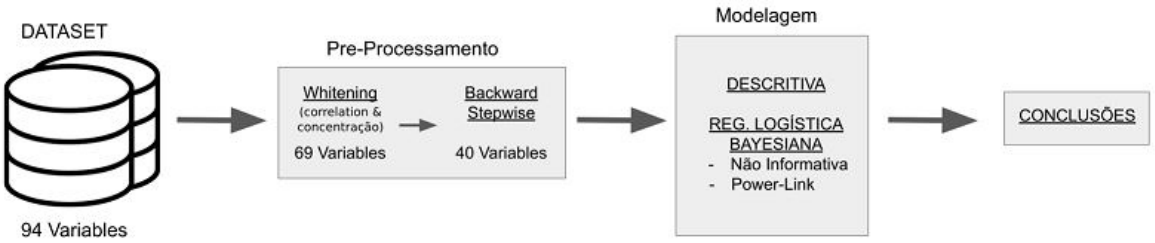

Figura 4 – Fluxo das análises desenvolvidas

Fonte: Elaborada pelo autor.

3.2.1 Estimação Clássica

Para a etapa de Modelagem, inicialmente considerou-se a abordagem Clássica, considerando o modelo de Regressão Logística com função de ligação logito e com os Estimadores de Máxima Verossimilhança para os parâmetros do modelo. Foi considerado o método de seleção de variáveis *BackWard* com o critério de nível de significância, considerado $\alpha = 5\%$, ou seja, inica-se a modelagem com todas as 69 variáveis provenientes da etapa de Pré-Modelagem, avalia-se o p-valor de cada variável, seleciona-se o maior p-valor, se este for superior à α, retira-se essa variável e treina-se novamente o modelo, realizando este procedimento até que nenhuma variável possua p-valor superior à α.

O modelo final é composto por 40 variáveis preditoras diferentes, todas significativas conisderando o nível de significância de 5%.

A Tabela 5 apresenta os tipos de variáveis preditoras presentes no modelo clássico proposto.Todas as variáveis cadastrais presentes na base de dados se mostraram significativas, 5 das 8 variáveis de consulta também entraram no modelo, o que evidencia a importância destas variáveis para explicar o evento de inadimplência. A maior parte das variáveis significativas para o modelo é do tipo "Débito", representando 45% do total de variáveis.

Tabela 5 – Tipos de Variáveis Preditoras - Modelo Clássico

Tipo	**Quantidade**	**Quantidade (%)**
Cadastral	3	7,5%
Consulta	5	12,5%
Débito	18	45,0%
Geográfica	14	35,0%
TOTAL	**40**	**100,0%**

Os critérios escolhidos para a avaliação de performance do modelo foram Kolmogorov-Smirnov (KS) e *Area Under the Curve* (AUC), que são bastante usuais para avaliação de modelos de crédito. O KS encontrado na Base de Desenvolvimento foi de **25,0%**, enquanto na Base de Validação, o KS encontrado foi **26,1%**. Em relação ao AUC, o resultado foi **33,1%** na Base de Desenvolvimento e **32,6%** na Base de Validação. Como os resultados são relativamente próximos, não há evidências de sobreajuste do modelo. Esses são os resultados finais considerando a abordagem Clássica. Após apresentar a análise descritiva das principais variáveis, o próximo passo é avaliar abordagens Bayesianas com o intuito de obter um resultado melhor em termos de KS e AUC, duas métricas que avaliam o poder discriminante do modelo, ou seja, no contexto deste estudo, o quão bem os modelos separam os bons e os maus pagadores.

3.2.2 Análise Descritiva - principais variáveis

A partir das 40 variáveis preditoras do modelo de Regressão Logística Clássico, selecionou-se as 10 variáveis que apresentaram maiores valores absolutos para os parâmetros do modelo, a fim de realizar uma análise descritiva mais profunda.

Os gráficos da Figura 5 trazem os gráficos de caixa (*boxplots*) para cada uma destas variáveis. Pode-se verificar uma grande quantidade de variáveis com valores bastante concentrados em um único valor, a exemplo da variável 'VP301_CP'. É importante destacar que, apesar da grande concentração em um único valor, essas variáveis possuem uma concentração inferior à 90%, critério utilizado como filtro na pré-modelagem. Outro ponto de destaque é que, entre as 10 variáveis com os maiores valores absolutos nos β's, 5 delas são variáveis com informações de débito ('VP165', 'VPDEB003','VPR','VP170' e 'VP158'), duas são variáveis com informações de consulta ('VP301_CP','VP308_CP'), duas com informações de geolocalização ('VPDCR0005', 'VPDCR0002') e uma variável cadastral('VP169').

Com o intuito de evidenciar a relação entre as principais variáveis preditoras e a variável resposta do modelo, os gráficos presentes nas Figura 6, apresentam duas faixas de valores para as variáveis, mostrando a volumetria de cada faixa e os patamares de inadimplência de cada uma. As faixas de valores foram determinadas avaliando as distribuições presentes nos gráficos de caixa de forma a procurar evidenciar diferentes níveis de inadimplência para um percentual do volume da Base de Dados .Por exemplo, para a variável 'VP165', tem-se uma concentração de 79% de registros abaixo ou igual ao valor zero, e este grupo possui uma inadimplência de 42,7%. Por outro lado, os 21% restantes, que possuem valores superiores a zero para essa variável, apresentam uma inadimplência de 62,2%, 19,5% acima do primeiro grupo. Esse efeito é observado com maior ou menor proporção para cada uma das variáveis do modelo.

Os gráficos de caixa para as demais 30 variáveis utilizadas nos modelos estão disponíveis nos Apêndicês.

3.2.3 Estimação Bayesiana - Prioris Não Informativas

A primeira aplicação de Inferência Bayesiana no estudo foi considerando o uso de prioris não-informativas para a estimação dos parâmetros da Regressão Logística. Para a realização desta etapa, considerou-se as 40 variáveis que foram selecionadas no modelo com a abordagem clássica.

Foram considerados alguns cenários utilizando para os parâmetros prioris com distribuição Normal ou Uniforme. Em todos os testes que serão apresentados foram utilizadas 10 mil iterações, a partir do algoritmo de MCMC. Além disso, com o intuito de reduzir a autocorrelação entre as amostra pseudo-aleatórias, a cada 10 amostras geradas, a décima era descartada ($Thinning = 10$).

Os testes considerando as prioris não-informativas descritas acima apresentaram patamares de KS próximos aos resultados encontrados considerando Inferência Clássica. O teste com melhor resultado em termos de KS é encontrado quando utiliza-se priori Uniforme $\beta \sim U(-1,1)$, que obtém **24,9%** de KS na Base de Desenvolvimento e **25,8%** na Base de Validação, com o AUC nas Bases de Desenvolvimento e Validação sendo **33,1%** e **32,7%**, respectivamente.

Avaliando esses resultados, observa-se patamares bastante próximos de KS e AUC dos modelos Clássico e Bayesianos, com alguma vantagem para o modelo Clássico. Porém, com a abordagem Bayesiana, os parâmetros do modelo passam a ser não apenas parâmetros mas também variáveis aleatórias com distribuição conhecida, o que é um ganho interessante para o acompanhamento do modelo, novos testes e estudos, além de mostrar uma vantagem na utilização de Inferência Bayesiana no contexto de Risco de Crédito. Contudo, a fim de alcançar melhores resultados também em termos de discriminação (KS e AUC), foram aplicadas outras abordagens, apresentadas na seção seguinte.

3.2.4 Estimação Bayesiana - Modelos Power Link

Nesta seção é proposta a utilização de *Power Links* como transformações da função logito para a modelagem. Com isso, utilizando prioris não-informativas Uniformes e Normais, foram testadas as aplicações da transformação *Power Link* e *Reversal Power Link*. Foram testados diferentes valores para o parâmetro λ, mantendo-o fixo e também realizado um teste considerando uma priori gamma $\lambda \sim \Gamma(1,1)$.

Os resultados do estudo considerando essa aplicação trouxeram ganhos no poder discriminante do modelo. Foram realizados mais de 60 testes, alterando o valor do parâmetro λ, quantidades de iterações, diferentes prioris com distribuições Normais e Uniformes, e as abordagens de *Power Link* e *Reversal Power Link*. Interessante notar que para os testes considerando 1.000 iterações, nenhum resultado convergiu para séries estacionárias, de acordo com o teste de Geweke.

O gráfico da Figura 7 apresenta os resultados de KS na base de validação para os melhores

testes utilizando *Power Links* e o resultado do modelo considerando a inferência clássica. Note que, apesar dos resultados estarem próximos, alguns testes apresentaram poder de discriminação superior ao modelo clássico. O modelo 1, teste com melhor performance em termos de KS, é o teste considerando uma priori Normal($\mu = 0$, σ^2 =10.000), com $\lambda = 0,2$ utilizando *Power Link* e com 10.000 iterações.

Já o gráfico da Figura 8 apresenta os resultados dos mesmos testes, porém agora considerando como métrica AUC. O modelo 1, com melhor resultado de KS, não apresenta o melhor resultado de AUC, ficando 0,1 abaixo até mesmo do modelo Clássico. Por outro lado, o modelo 12, que considera também uma priori Normal($\mu = 0$, σ^2 =10.000), $\lambda = 0,2$ e 10.000 iterações, porém utilizando *Reversal Power Link*, apresentou o melhor resultado em termos desta métrica, sendo superior também ao modelo clássico.

A Tabela 6 traz os parâmetros utilizados em cada um dos testes apresentados nas Figuras 7 e 8. Vale destacar que as prioris apresentadas se referem sempre às distribuições Uniforme e Normal. Além disso, esses são os resultados apenas para os testes que apresentaram estacionaridade de acordo com o teste de Geweke. Foram testadas diversas combinações de prioris, diferentes valores para λ, *Power Links* e quantidade de iterações, porém grande parte dos testes com iterações menores que 10.000 não convergiram para séries estacionárias no momento da estimação dos parâmetros. Sendo assim, esses resultados não foram considerados para as análises finais. Entre todos os testes realizados que convergiram apenas um, o modelo 13 apresentou resultados considerando uma priori também para o parâmetro λ. Os demais testes consideraram valores arbitrario para λ, variando de 0,2 até 20. Com relação ao número de iterações para cada teste, cada combinação dos demais parâmetros foi testada com 1.000, 5.000 e 10.000 iterações.

Tabela 6 – Descrição dos testes apresentados

Teste	**Abreviação**	**Priori**	**Transformação na F. de Ligação**	λ	**Número de Iterações**	**KS (Validação)**	**AUC (Validação)**
modelo 1	mod1	Normal(0,10.000)	*Power Link*	0,2	10.000	26,8%	32,5%
modelo 2	mod2	Normal(0,10.000)	*Power Link*	0,2	5.000	26,6%	32,5%
modelo 3	mod3	Normal(0,10.000)	*Power Link*	0,4	5.000	26,5%	32,5%
modelo 4	mod4	Normal(0,10.000)	*Power Link*	0,4	10.000	26,5%	32,5%
modelo 5	mod5	Uniforme(-1,1)	*Reversal Power Link*	0,2	10.000	26,4%	32,5%
modelo 6	mod6	Uniforme(-1,1)	*Power Link*	0,6	10.000	26,2%	32,6%
modelo 7	mod7	Normal(0,10.000)	*Power Link*	0,6	5.000	26,3%	32,6%
modelo 8	mod8	Normal(0,10.000)	*Power Link*	0,8	10.000	26,3%	32,6%
Modelo Clássico	Mod. Classico	modelo Classico	-	-	-	26,1%	32,6%
modelo 10	mod10	Uniforme(-1,1)	*Reversal Power Link*	1	10.000	26,0%	32,7%
modelo 11	mod11	Uniforme(-1,1)	*Reversal Power Link*	0,6	10.000	25,8%	32,7%
modelo 12	mod12	Normal(0,10.000)	*Reversal Power Link*	0,2	10.000	25,5%	33,0%
modelo 13	mod13	Normal(0,10.000)	*Power Link*	$\Gamma(1,1)$	10.000	26,0%	32,7%
modelo 14	mod14	Normal(0,100)	-	-	10.000	22,1%	32,8%
modelo 15	mod15	Uniforme(-1,1)	-	-	10.000	25,8%	32,7%
modelo 16	mod16	Uniforme(-2,2)	-	-	10.000	19,9%	32,7%

Selecionando o modelo 1, teste com priori Normal($\mu = 0$, σ^2 =10.000) e transformação *Power Link* com λ = 0,2, que apresentou melhor resultado em termos de KS, avaliou-se mais a fundo os intervalos de credibilidade para a mediana de cada parâmetro associado às variáveis preditoras do modelo, apresentados no gráfico da Figura 9. O gráfico traz como limites inferior

e superior os percentis 2,5% e 97,5%, com preenchimento mais forte no valor que representa a mediana das estimativas. Note que, para esse teste, os intervalos de nenhuma das variáveis possui o valor de zero.

A Figura 10 traz a mesma análise de Intervalo de Credibilidade, mas agora para o modelo 12, teste com priori Normal($\mu = 0$, σ^2 =10.000) e transformação *Reversal Power Link* com λ = 0,2, que apresentou melhor resultado em termos de AUC. Note que para este teste, alguns dos parâmetros possuem Intervalos de Credibilidade que contém o valor zero, diferente do teste com o maior KS. O parâmetro β associado a 16ª variável (VPCON0001 - uma variável de consulta) assim como no modelo da Figura 9 apresenta grande dispersão.

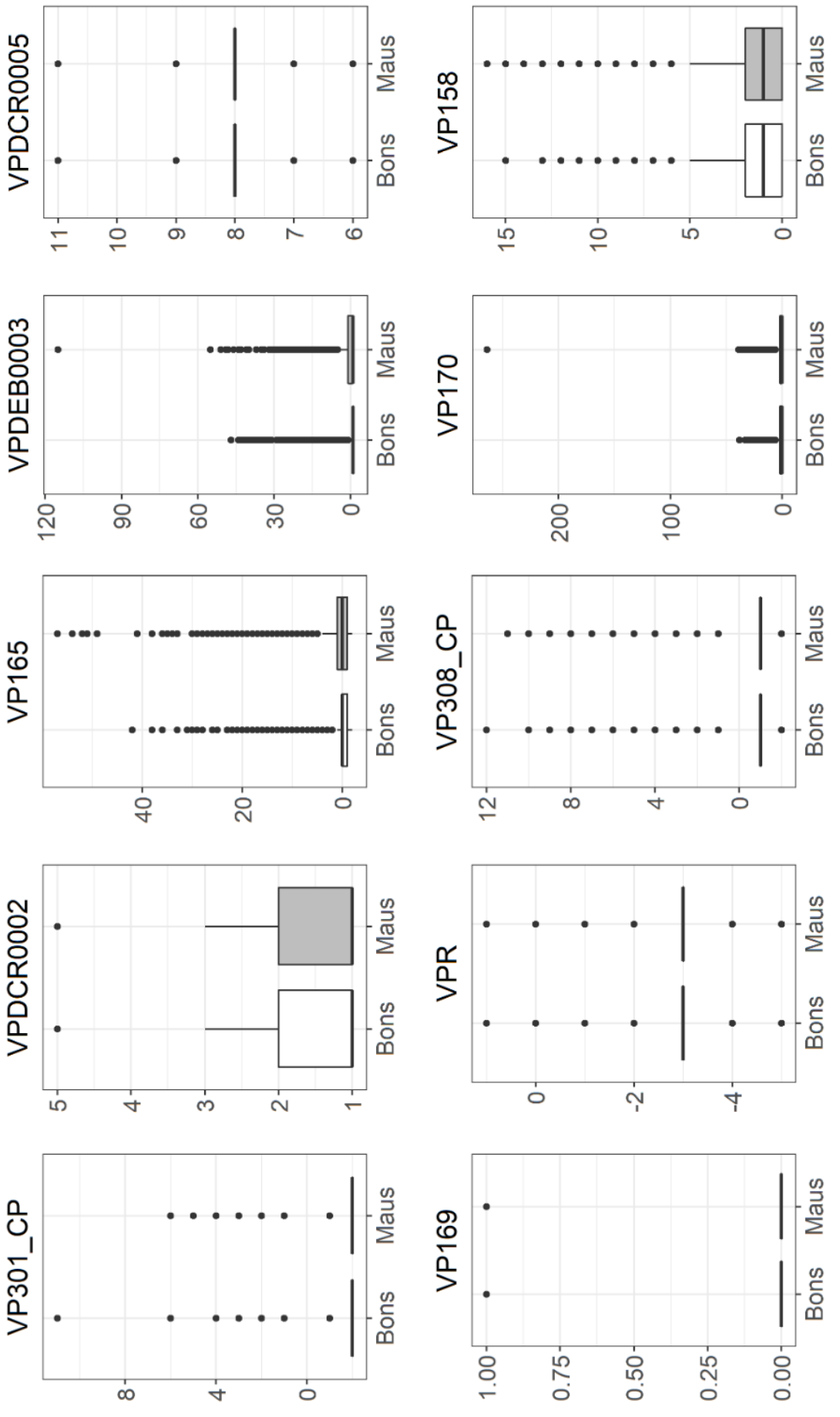

Figura 5 – *Boxplots* variáveis com maiores β's

Fonte: Elaborada pelo autor.

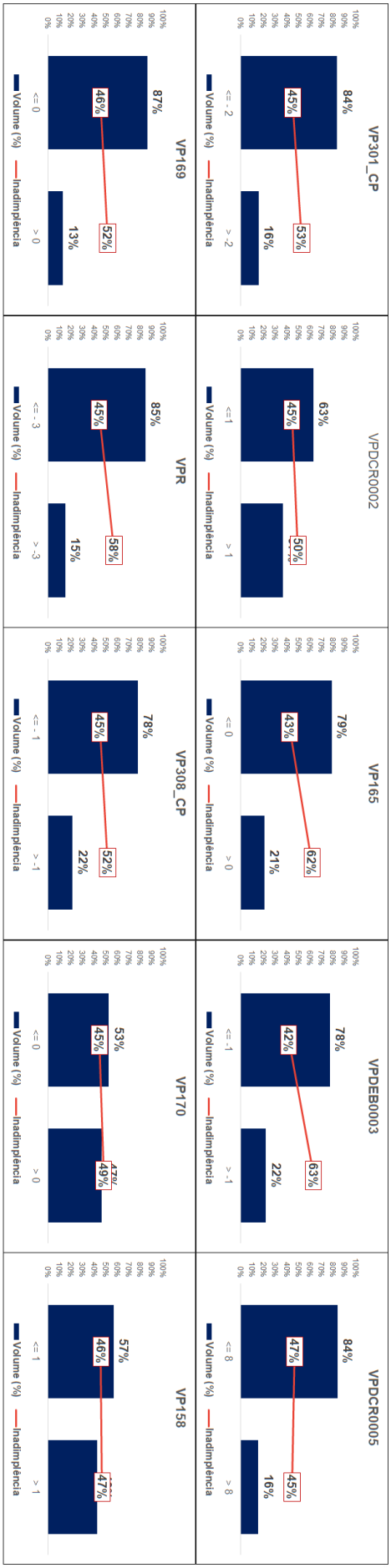

Figura 6 – Volume e Inadimplência

Fonte: Elaborada pelo autor.

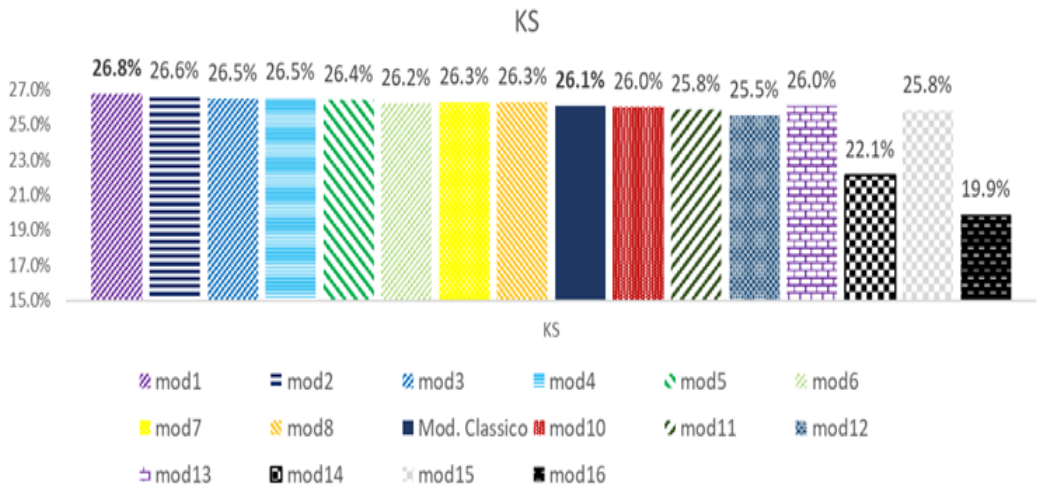

Figura 7 – KS dos modelos testados

Fonte: Elaborada pelo autor.

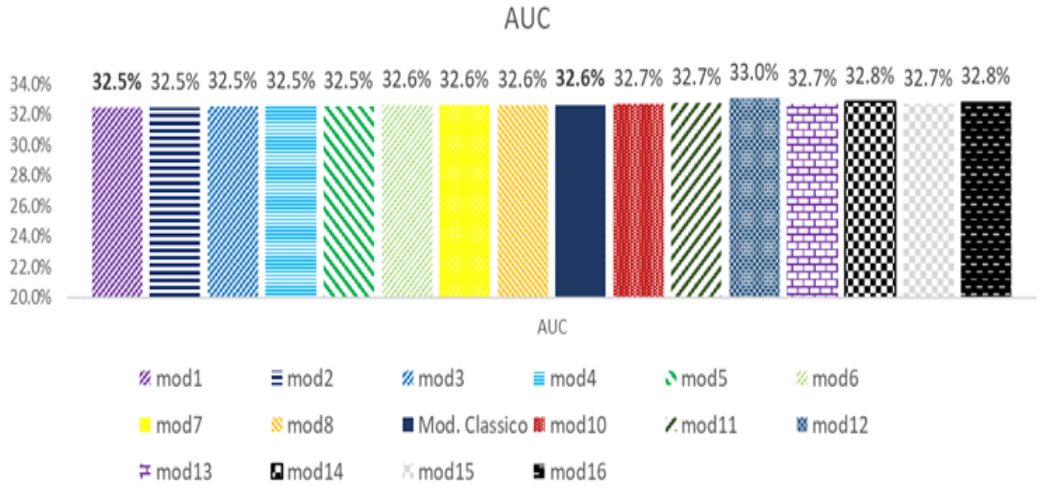

Figura 8 – AUC dos modelos testados

Fonte: Elaborada pelo autor.

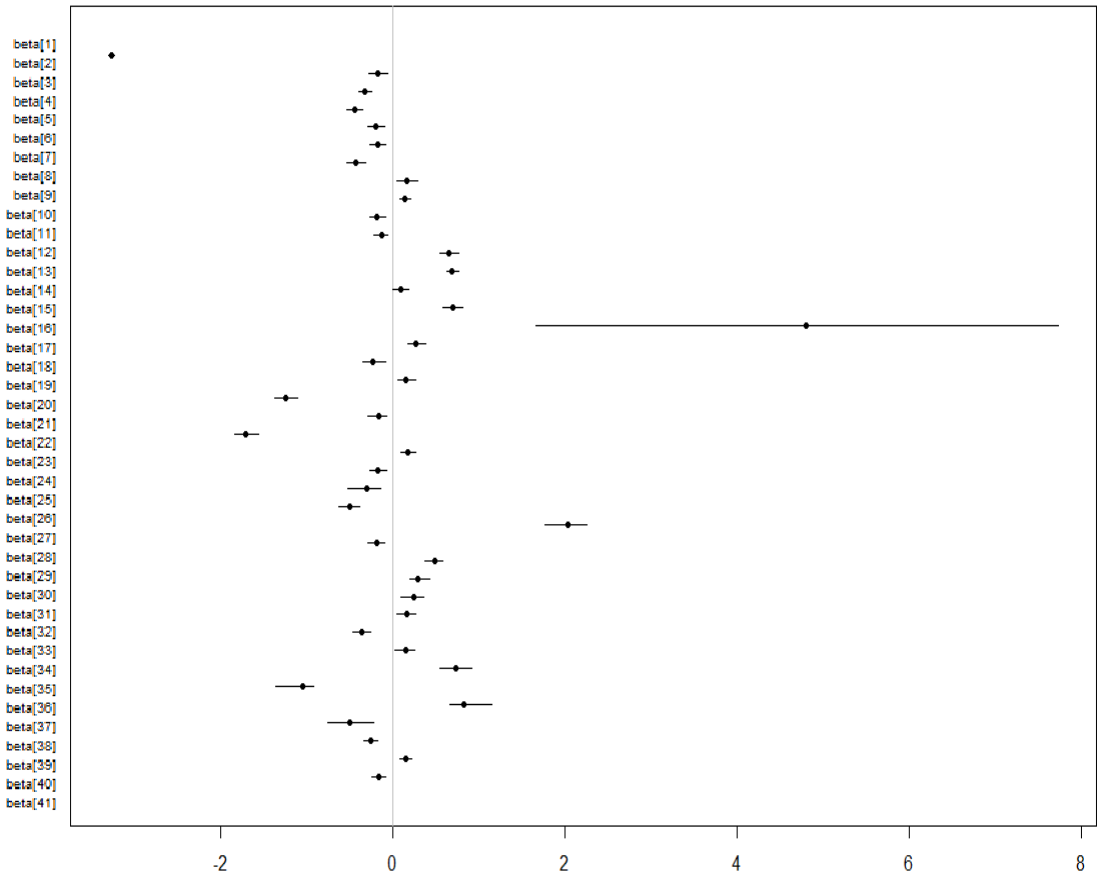

Figura 9 – Intervalos de Credibilidade para cada parâmetro (Considerando o modelo 1).

Fonte: Elaborada pelo autor.

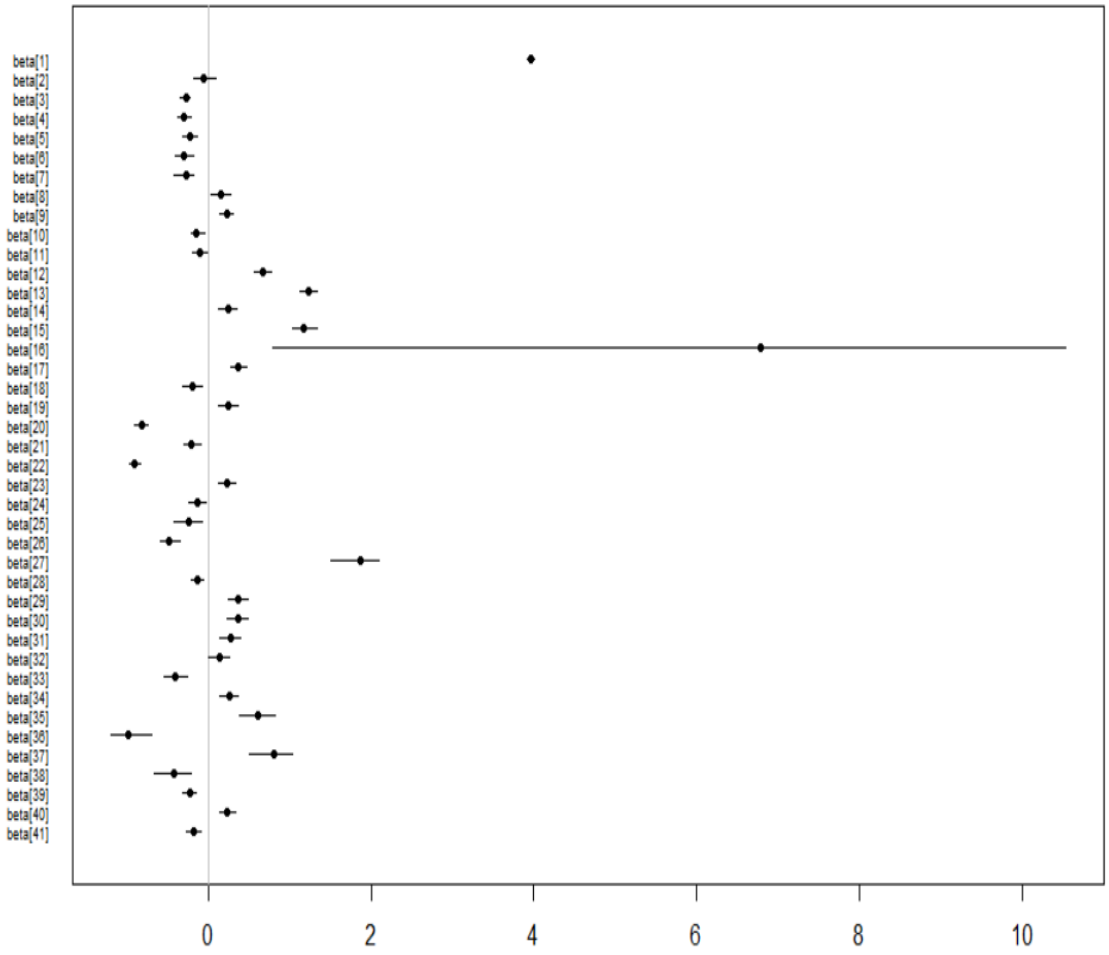

Figura 10 – Intervalos de Credibilidade para cada parâmetro (Considerando o modelo 12).

Fonte: Elaborada pelo autor.

CAPÍTULO

4

CONCLUSÕES

A aplicação das técnicas Bayesianas em modelos de risco de crédito apresentaram resultados bastante próximos aos da inferência clássica em termos de discriminação de bons e maus pagadores. Foi possível obter modelos com resultados de KS levemente superiores e outros com resultados de AUC levemente superiores, sempre comparando com o modelo clássico. Esses resultados são razoáveis: o fato de incorporar uma priori aos parâmetros do modelo não necessariamente deve implicar em um ganho em discriminação, mas ainda assim foi possível obter bons resultados. Sendo assim, mesmo se cosiderar que o patamar de discriminação dos modelos foi mantido com os modelos Bayesianos propostos, agora cada parâmetro do modelo possui uma distribuição própria e conhecida, o que traz ganhos no acompanhamento do modelo e das variáveis, possibilitando definir intervalos de credibilidade Bayesianos, e realizar demais pesquisas e estudos. Portanto, pode-se concluir que não se deve descartar a utilização de modelos baysianos no contexto de risco de crédito.

Outros estudos relacionados podem ser desenvolvidos a partir dos testes apresentados nesta dissertação, como a utilização de *Power Priors*, onde utiliza-se o conhecimento prévio de um estudo anterior como priori informativa (os resultados encontrados nesta dissertação poderiam ser utilizados como conhecimento prévio acerca do tema), outras prioris, informativas ou não, poderiam ser avaliadas também. Além disso, a aplicação de diferentes transformações *Power Link* na função de ligação da Regressão Logística e a aplicação de outras técnicas de modelagem também podem ser avaliadas em estudos futuros. Essa dissertação teve como intuito desafiar a técnica usualmente aplicada no mercado, por isso apenas modelos de Regressão Logística foram avaliados.

Outro tema de bastante impacto no contexto de risco de crédito é o da inferência de performance do público que teve a proposta de crédito negada. O uso de técnicas Bayesianas para a realização desta inferência pode ser bastante relevante e trazer resultados que gerem ganhos nas operações de crédito das instituições financeiras, e que possibilite que um outro perfil de

público passe a ter acesso a crédito que antes seria negado, sendo também uma oportunidade para os consumidores.

REFERÊNCIAS

ATIYA, A. **Bankruptcy prediction for credit risk using neural networks: A survey and new results**. IEEE Transactions on Neural Networks, 2001. ISBN 9781417642595. Disponível em: <https://ieeexplore.ieee.org/document/935101>. Citado na página 23.

BAZáN, J. L.; TORRES-AVILéS, F.; SUZUKI, A. K.; LOUZADA, F. **Power and reversal power links for binary regressions: An application for motor insurance policyholders**. Applied Stochastic Models in Business and Industry, 33(1), 22-34, 2017. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/asmb.2215>. Citado na página 31.

BCB. **Relatório de Economia Bancária e Crédito**. BANCO CENTRAL DO BRASIL, Brasília, DF, 2020. Disponível em: <https://www.bcb.gov.br/publicacoes/relatorioeconomiabancaria>. Citado nas páginas 21 e 22.

BORGES, L. C. **Análise bayesiana do modelo fatorial dinâmico para um vetor de séries temporais usando distribuições elípticas**. Tese (Doutorado), 2008. Disponível em: <https://www.teses.usp.br/teses/disponiveis/45/45133/tde-11092008-143337/publico/tese_liviaborges.pdf>. Citado na página 32.

CHEN, G.; ÅSTEBRO, T. **Bound and collapse Bayesian reject inference for credit scoring**. Journal of the Operational Research Society volume 63, 1374–1387 (2012), 2011. ISBN 9781417642595. Disponível em: <https://link.springer.com/article/10.1057/jors.2011.149>. Citado na página 24.

DIEESE. **A evolução do crédito na economia brasileira 2008-2013**. DIEESE - Departamento Intersindical de Estatística e Estudos Socioeconômicos, SP, 2014. Disponível em: <https://bancariosdf.com.br/portal/wp-content/uploads/2014/08/notaTec135Credito.pdf>. Citado na página 21.

GELMAN, A.; CARLIN, J. B.; STERN, H. S.; RUBIN, D. B. **Bayesian data analysis (Third edition)**. [S.l.]: Chapman and Hall/CRC, 1995. Citado nas páginas 28 e 29.

KING G., . Z. L. **Logistic regression in rare events data**. Political analysis, 9(2), 137-163 (2001), 2001. ISBN 9781417642595. Disponível em: <https://www.cambridge.org/core/services/aop-cambridge-core/content/view/1E09F0F36F89DF12A823130FDF0DA462/S1047198700003740a.pdf/logistic-regression-in-rare-events-data.pdf>. Citado na página 30.

KOLáCEK J., . R. M. **Assessment of scoring models using information value**. 19, 2010. ISBN 9781417642595. Disponível em: <https://www.researchgate.net/publication/258421575_Assessment_of_scoring_models_using_information_value>. Citado na página 33.

LEONG, C. K. **Credit risk scoring with bayesian network models. Computational Economics**. Computational Economics, 47(3), 423-446 (2016), 2016. ISBN 9781417642595. Disponível em: <https://link.springer.com/article/10.1007/s10614-015-9505-8>. Citado na página 24.

SCHERVISH, M. J.; DEGROOT, M. H. **Probability and statistics**. [S.l.]: Pearson Education, 2014. ISBN 100321500466. Citado na página 25.

WANG, S. W. Y.; LAI, K. K. **A new fuzzy support vector machine to evaluate credit risk**. IIEEE Transactions on Fuzzy Systems, 2005. ISBN 9781417642595. Disponível em: <https://ieeexplore.ieee.org/document/1556587>. Citado na página 23.

WILHELMSEN M., D. X. K. H. T. . F. M. **Bayesian modelling of credit risk using integrated nested laplace approximations**. NR publication, 1-25 (2009), 2009. ISBN 9781417642595. Disponível em: <https://link.springer.com/article/10.1007/s10614-015-9505-8>. Citado na página 24.

XIA, Y. **A Novel Reject Inference Model Using Outlier Detection and Gradient Boosting Technique in Peer-to-Peer Lending**. IEEE Access, 2019. ISBN 9781417642595. Disponível em: <https://ieeexplore.ieee.org/document/8758218>. Citado na página 23.

ZHANG, X.; YANG, Y.; ZHOU, Z. **A novel credit scoring model based on optimized random forest**. IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), 2018. ISBN 9781417642595. Disponível em: <https://ieeexplore.ieee.org/document/8301707>. Citado na página 23.

ZHANG, X.; ZHOU, Z. **Credit Scoring Model based on Kernel Density Estimation and Support Vector Machine for Group Feature Selection**. International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2018. ISBN 9781417642595. Disponível em: <https://ieeexplore.ieee.org/document/8554524>. Citado na página 23.

CAPÍTULO

5

APÊNDICES

5.1 *Boxplot* demais variáveis

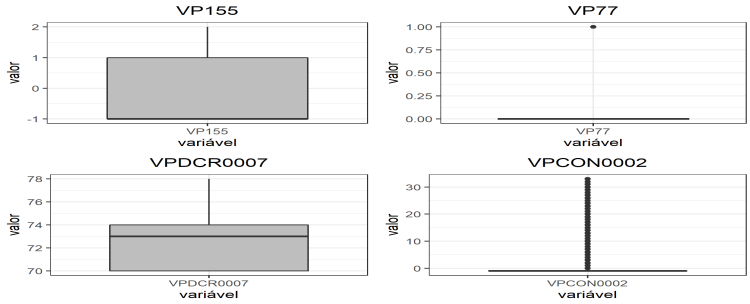

Figura 11 – Boxplot demais variáveis - parte 1.

Fonte: Elaborada pelo autor.

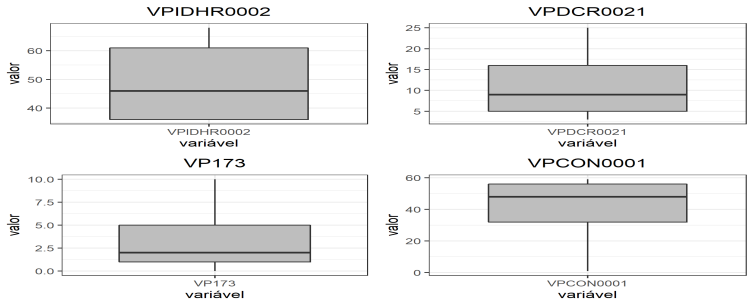

Figura 12 – Boxplot demais variáveis - parte 2.

Fonte: Elaborada pelo autor.

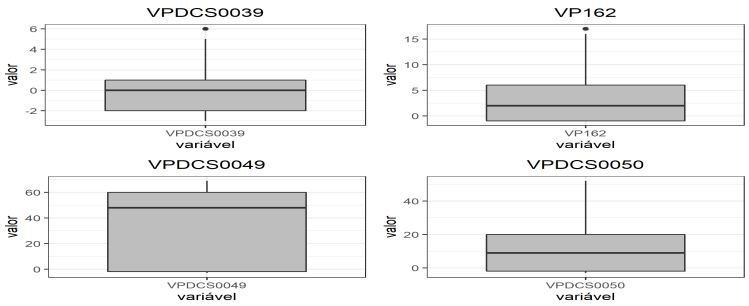

Figura 13 – Boxplot demais variáveis - parte 3.

Fonte: Elaborada pelo autor.

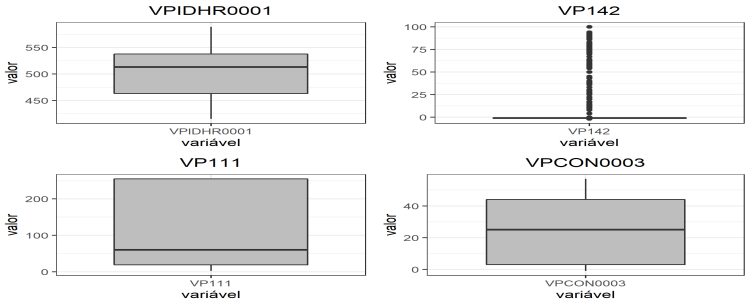

Figura 14 – Boxplot demais variáveis - parte 4.

Fonte: Elaborada pelo autor.

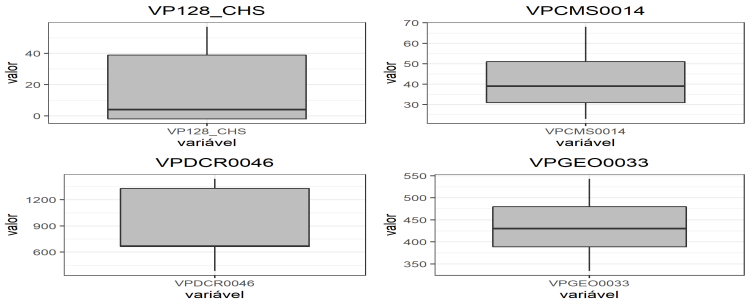

Figura 15 – Boxplot demais variáveis - parte 5.

Fonte: Elaborada pelo autor.

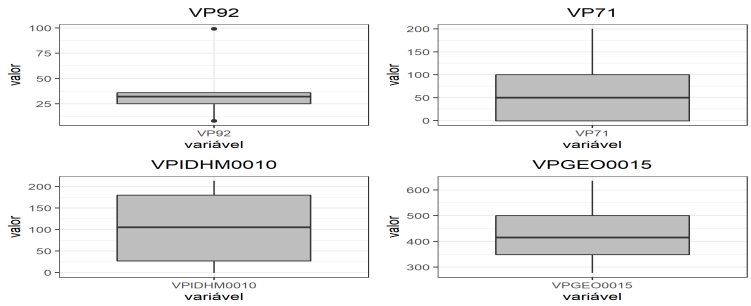

Figura 16 – Boxplot demais variáveis - parte 6.

Fonte: Elaborada pelo autor.

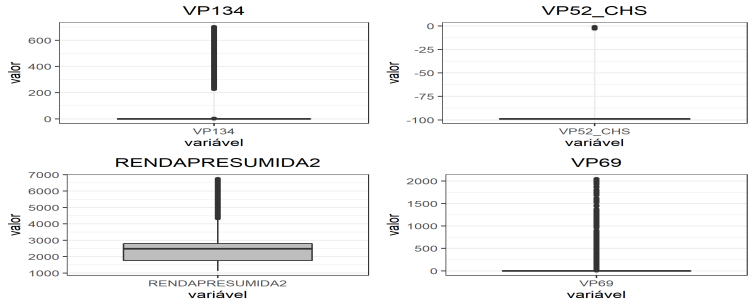

Figura 17 – Boxplot demais variáveis - parte 7.

Fonte: Elaborada pelo autor.

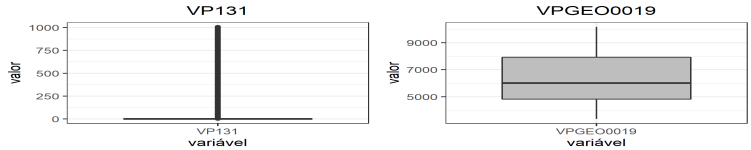

Figura 18 – Boxplot demais variáveis - parte 8.

Fonte: Elaborada pelo autor.

5.2 Códigos em R

```
1:
2:
3:
4: #-------------------------#
5: #IMPORTANDO A BASE DE DADOS#
6: #-------------------------#
7:
8:
9: setwd("C:/Users/Erick/Google Drive/USP/2020/Pesquisa")
10:
11: source("libs_e_funcoes.r")
12:
13: #Base de dados utilizada: 99.766 registros com 105 variaveis -
      essa virou a Base_MECAI_v1_bkp
14: #Base_MECAI.TXT: apenas variáveis a serem utilizadas, com 97
      vars
15: dados <- fread("Base_MECAI.csv")
16:
17: head(dados)
18:
19: #-------------------------#
20: #Pré-Seleção das Variáveis#
21: #-------------------------#
22:
23: #filtrando variáveis com mais de 90% de volume E com correlação
       >70%
24:
25:
26: #FILTRO DE VOLUME (90%)#
27:
28: dados_vps = as.data.table(select (dados,-c(ID,
29:                                             Ind_Bom_Merc,
30:                                             IND_BOM,
31:                                             VPR_VLRM)))
32:
33:
34: lista_vps = names(dados_vps)
35: lista_volume = NULL
36: lista_filtro_volume = NULL
37: for( i in 1:length(lista_vps)){
```

```
38:   #lista apenas dos as vps filtadadas por alto vol:
39:   lista_filtro_volume = calcula_volume(dados,lista_vps[i],0.90)
40:   #lista compleata com o vol. maximo em um valor
41:   lista_volume = calcula_volume_v2(dados,lista_vps[i])
42:
43: }
44:
45: lista_filtro_volume = as.data.table(lista_filtro_volume)
46: lista_volume = as.data.table(lista_volume)
47:
48: lista_filtro_volume = lista_filtro_volume[order(-vol_max)]
49: lista_volume = lista_volume[order(-vol_max)]
50:
51: lista_filtro_volume
52: lista_volume
53: lista_filtro_volume$V1
54: #TIRANDO AS VARS COM ALTA CONCENTRAÇÃO DE VOLUME
55: dados_vps = as.data.table(select (dados_vps,-
      lista_filtro_volume$V1))
56: #variáveis filtradas por volume:
57: #V1 vol_max
58: #1: VPCCF0007    0.95
59: #2:      VP39    0.94
60: #3: VPCHS0001    0.94
61: #4:      VP45    0.93
62: #5:      VP49    0.93
63: #6:      VP52    0.93
64: #7: VPCCF0006    0.93
65: #8:     VP140    0.93
66: #9: VPDCR0031    0.93
67: #10:  VP298_CP    0.92
68: #11:      VP65    0.92
69: #12: VPCCF0009     0.9
70:
71: #salvando a volumetria maxima encontrada para cada var
72: write.csv(lista_volume,"lista_volume.csv",row.names = F)
73:
74: correlations <- cor(dados_vps)
75: corrplot(correlations, method="circle")
76:
77: #(correlations > 0.7 & correlations < 1)
78:
```

```
#vou ver as corrs no excel:
write.csv(correlations,"correlação_vps.csv",row.names = F)

#apos o filtro de volume (90%) foi verificada a correlação entre as variáveis.

vars_correlacionadas <- as.data.table(select (dados_vps,c('VP128_CHS','VP013_018r','VP013_014r',
                                                          'VP001_006r','VPIDHM0010','VPIDHM0009','VPIDHM0008','VPDCS0049','VPDCS0027','VPDCS0007','VPDCM0038','VPDCM0035','VPDCM0027','VP160','VP121','VP12','VP111','VP128')))

corr_vars_corr <- cor(vars_correlacionadas)

#para manter apenas a parte superior no gráfico:
for(i in 1:dim(corr_vars_corr)[1]){
  for(j in 1:dim(corr_vars_corr)[1]){
    if(i>j){
      corr_vars_corr[i,j] = 0
    }
  }
}
corrplot(corr_vars_corr, method="circle")
aux = corr_vars_corr


row.names(corr_vars_corr)

row.names(aux) = c("VPF2_0001","VPF2_0002","VPF2_0003","VPF2_0004","VPF2_0005","VPF2_0006",
                   "VPF2_0007","VPF2_0008","VPF2_0009","VPF2_0010","VPF2_0011","VPF2_0012",
                   "VPF2_0013","VPF2_0014","VPF2_0015","VPF2_0016","VPF2_0017","VPF2_0018")

colnames(aux) = c("VPF2_0001","VPF2_0002","VPF2_0003","VPF2_0004","VPF2_0005","VPF2_0006",
                  "VPF2_0007","VPF2_0008","VPF2_0009","VPF2_0010","VPF2_0011","VPF2_0012",
                  "VPF2_0013","VPF2_0014","VPF2_0015","VPF2_0016","VPF2_0017","VPF2_0018")
```

```
110:
111: corrplot(aux, method="circle")
112:
113:
114: vars_correlacionadas_v2 <- as.data.table(select (
       vars_correlacionadas,-c("VP128","VP013_014r","VP001_006r","
       VPDCM0035","VPDCM0027","VPIDHM0009","VPIDHM0008","VPDCM0038
       ","VPDCS0027","VPDCS0007","VP160","VP12","VP121")))
115:
116:
117:
118: corr_vars_corr_v2 <- cor(vars_correlacionadas_v2)
119:
120: #para manter apenas a parte superior no gráfico:
121: for(i in 1:dim(corr_vars_corr_v2)[1]){
122:   for(j in 1:dim(corr_vars_corr_v2)[1]){
123:     if(i>j){
124:       corr_vars_corr_v2[i,j] = 0
125:     }
126:     if(abs(corr_vars_corr_v2[i,j]) < 0.65){
127:       #para facilitar quando eu vejo no gráfico
128:       corr_vars_corr_v2[i,j] = 0
129:     }
130:   }
131: }
132: corrplot(corr_vars_corr_v2, method="circle")
133: #APENAS VPS NÃO CORRELACIONADAS
134:
135: aux = corr_vars_corr_v2
136: #aux é apenas para a dissertação - para omitir o nome das variá
       veis
137:
138: row.names(aux) = c("VPF2_0001","VPF2_0002","VPF2_0005","
       VPF2_0008","VPF2_0017")
139: colnames(aux) = c("VPF2_0001","VPF2_0002","VPF2_0005","
       VPF2_0008","VPF2_0017")
140: corrplot(aux, method="circle")
141:
142: #lista_volume = as.data.table(lista_volume)
143: #lista_volume[(V1 =="VP121")|(V1 =="VP111")|(V1 =="VP12"),]
144:
145: #mantidas:
```

```
146: #VP128_CHS
147: #VP013_018r
148: #VPIDHM0010
149: #VPDCS0027
150: #VPDCS0049
151: #VP111
152:
153: #dropar por corr e maior concentração de volume:
154: #VP128
155: #VP013_014r
156: #VP001_006r
157: #VPDCM0035
158: #VPDCM0027
159: #VPIDHM0009
160: #VPIDHM0008
161: #VPDCM0038
162: #VPDCS0007
163: #VP160
164: #VP12
165: #VP121
166:
167:
168: setwd("C:/Users/Erick/Google Drive/USP/2020/Pesquisa")
169:
170: source("libs_e_funcoes.r")
171:
172:
173: dados <- fread("FILTRADA.csv")
174:
175:
176: names(dados)
177:
178: #CRIANDO BASE E RODANDO MODELO CLASSICO #
179:
180: #base de dados apenas com a variavel target + as vars que serão
        consideradas para a modelagem
181:
182: dados_resumo <- as.data.table(select(dados,-c("Ind_Bom_Merc","
       ID","RANDOM")))
183:
184: dados_resumo[,.N,by=FLAG_DESENV]
185: #FLAG_DESENV     N
```

```
#1:            1 69930
#2:            0 29836

dados_resumo[,.N,by=IND_BOM]


correlations <- cor(dados_resumo)
corrplot(correlations, method="circle")


treino = dados_resumo[FLAG_DESENV ==1,]
teste = dados_resumo[FLAG_DESENV ==0,]
?glm



treino[,.N,by=IND_BOM]
#INAD TREINO: 0.4679394
teste[,.N,by=IND_BOM]
#INAD TESTE: 0.4660142


treino<- select(treino,-c("FLAG_DESENV"))


modelo.fit = glm(formula =IND_BOM~. ,data=treino, family =
    binomial)


treino<- select(treino,-c('VPDCR0008','VPDCR0014','VPDCR0016','
    VPDCR0024','VPDCR0036'))

modelo.fit = glm(formula =IND_BOM~. ,data=treino, family =
    binomial)

summary.glm(modelo.fit)

write.csv2(summary.glm(modelo.fit)$coefficients,"summary_modelo
    .csv")



```

```
224: ########################
225: #modelo Frequentista #
226: ########################
227: #Apenas vars do backward:
228: modelo.fit = glm(formula =IND_BOM~RENDAPRESUMIDA2+VPCMS0014+
       VP301_CP+VP308_CP+VP158+VP170+VP173+
229:                                     VP69+VP77+VP92+VPCON0001+
       VPCON0002+VPCON0003+VP111+VP131+VP142+
230:                                     VP155+VP162+VP165+VP71+
       VPDEB0003+VP134+VP169+VPDCR0007+VPDCR0021+
231:                                     VPDCR0046+VPDCS0039+VPDCS0049
       +VPDCS0050+VPGE00015+VPGE00019+
232:                                     VPGE00033+VPIDHM0010+
       VPIDHR0001+VPIDHR0002+VPDCR0002+VPDCR0005+
233:                                     VPR+VP52_CHS+VP128_CHS
234:                    ,data=treino, family = binomial)
235:
236:
237:
238: summary.glm(modelo.fit)
239:
240: write.csv2(summary.glm(modelo.fit)$coefficients,"
       summary_backward_glm.csv")
241:
242:
243: treino[,prob_bom := predict(modelo.fit,newdata=treino,type = "
       response")]
244:
245: teste[,prob_bom := predict(modelo.fit,newdata=teste,type = "
       response")]
246:
247: KS(treino$prob_bom,treino[,.(IND_BOM)])
248: #0.2496049
249:
250: KS(teste$prob_bom,teste[,.(IND_BOM)])
251: #0.2605624
252:
253:
254:
255:
256:
257:
```

```
setwd("C:/Users/Erick/Google Drive/USP/2020/Pesquisa")


source("libs_e_funcoes.r")


dados <- fread("FILTRADA.csv")

dim(dados)

dados[,constante := rep(1,dim(dados)[1])]

dados_amostra <- dados[RANDOM <= 0.01,]

dim(dados_amostra)
#[1] 314 322

N <- dim(dados_amostra)[1]
y <- dados_amostra %>% select(IND_BOM) %>% pull()



X_FULL <- as.matrix.data.frame(dados[,.(constante,
    RENDAPRESUMIDA2,VPCMS0014,VP301_CP,VP308_CP,VP158,VP170,
    VP173,VP69,VP77,VP92,VPCON0001,VPCON0002,VPCON0003,VP111,
    VP131,VP142,VP155,VP162,VP165,VP71,VPDEB0003,VP134,VP169,
    VPDCR0007,VPDCR0021,VPDCR0046,VPDCS0039,VPDCS0049,VPDCS0050,
    VPGEO0015,VPGEO0019,VPGEO0033,VPIDHM0010,VPIDHR0001,
    VPIDHR0002,VPDCR0002,VPDCR0005,VPR,VP52_CHS,VP128_CHS)])



y_valid <- dados[FLAG_DESENV == 0,] %>% select(IND_BOM) %>%
    pull()



X_valid <- as.matrix.data.frame(dados[FLAG_DESENV == 0,.(
    constante,RENDAPRESUMIDA2,VPCMS0014,VP301_CP,VP308_CP,VP158
                                                          ,VP170
    ,VP173,
```

```
291:                                                                VP69,
      VP77,VP92,VPCON0001,VPCON0002,VPCON0003,VP111,VP131,VP142,
292:                                                                VP155,
      VP162,VP165,VP71,VPDEB0003,VP134,VP169,VPDCR0007,VPDCR0021,
293:
      VPDCR0046,VPDCS0039,VPDCS0049,VPDCS0050,VPGEO0015,VPGEO0019,
294:
      VPGEO0033,VPIDHM0010,VPIDHR0001,VPIDHR0002,VPDCR0002,
      VPDCR0005,
295:                                                                VPR,
      VP52_CHS,VP128_CHS
296: )])
297:
298:
299: dados <- fread("FILTRADA.csv")
300:
301: dim(dados)
302:
303: dados[,constante := rep(1,dim(dados)[1])]
304:
305:
306: X_FULL <- as.matrix.data.frame(dados[,.(constante,
      RENDAPRESUMIDA2,VPCMS0014,VP301_CP,VP308_CP,VP158,VP170,
      VP173,VP69,VP77,VP92,VPCON0001,VPCON0002,VPCON0003,VP111,
      VP131,VP142,VP155,VP162,VP165,VP71,VPDEB0003,VP134,VP169,
      VPDCR0007,VPDCR0021,VPDCR0046,VPDCS0039,VPDCS0049,VPDCS0050,
      VPGEO0015,VPGEO0019,VPGEO0033,VPIDHM0010,VPIDHR0001,
      VPIDHR0002,VPDCR0002,VPDCR0005,VPR,VP52_CHS,VP128_CHS)])
307:
308:
309:
310:
311: y_valid <- dados[FLAG_DESENV == 0,] %>% select(IND_BOM) %>%
      pull()
312:
313:
314:
315: X_valid <- as.matrix.data.frame(dados[FLAG_DESENV == 0,.(
      constante,RENDAPRESUMIDA2,VPCMS0014,VP301_CP,VP308_CP,VP158,
      VP170,VP173,VP69,VP77,VP92,VPCON0001,VPCON0002,VPCON0003,
      VP111,VP131,VP142,VP155,VP162,VP165,VP71,VPDEB0003,VP134,
      VP169,VPDCR0007,VPDCR0021,VPDCR0046,VPDCS0039,VPDCS0049,
```

```
      VPDCS0050,VPGEO0015,VPGEO0019,V0PGEO0033,VPIDHM0010,
      VPIDHR0001,VPIDHR0002,VPDCR0002,VPDCR0005,VPR,VP52_CHS,
      VP128_CHS)])
316:
317:
318: ########################
319: #CRIANDO CADA PROB  E#
320: #CALCULANDO KS      #
321: ######################
322:
323: Modelos  <- c(ModelNorm_pl,ModelNorm_rvpl,ModelUniform_pl,
      ModelUniform_rvpl)
324: #Mod 1: Modelo com priori Normal(0,10000) com PL;
325: #Mod 2: Modelo com priori Normal(0,10000) com Reversal PL;
326: #Mod 3: Modelo com Uniforme(-1,1) com PL;
327: #Mod 4: Modelo com Uniforme(-1,1) com Reversal PL;
328:
329:
330:
331: interacoes = c(5000,10000)
332: resumo_result = NULL
333: Modelos = c(1,2)
334:
335: aux_tempo = 1/(length(interacoes)*length(Modelos))#*length(
      lambdas) - sum(lambdas<0.5))
336: ind_tempo = 0
337:
338:
339: base_prob_perf = fread("C:/Users/Erick/Google Drive/USP/2020/
      Pesquisa/PowerLinks_LaplacesDemon/
      base_prob_perfl_LD_POWERLINK.csv")
340:
341: resumo_result = fread(
342:           "C:/Users/Erick/Google Drive/USP/2020/Pesquisa/
      PowerLinks_LaplacesDemon/resumo_result_PL.csv")
343:
344: J = 41
345: # ENTAO SERAO 2 FOR'S: UM COM LAMBDA > 0.5 E UM PARA LAMBDA <
      0.5, SEM O TESTE COM MODELO I = 3!
346: for (it in interacoes){
347:   for(i in Modelos){
```

```
348:       Fit <- readRDS(paste0("C:/Users/Erick/Google Drive/USP
     /2020/Pesquisa/PowerLinks_LaplacesDemon/",
349:                             "powerLinks_v3/fitLD_mod",i,"
     _lambda_priori_it",it,".rds"))
350:       if(i %in% c(2)){
351:         rev = TRUE #modelo com reversal PL
352:       }else{
353:         rev = FALSE
354:       }
355:
356:       prob = prob_power_link(X_FULL,Fit,1,rev) #para o
     calculo de KS a transformação aqui não impacta!
357:
358:       base_prob_perf[,paste0("fitLD_mod",i,"_lambda_priori_it
     ",it):=prob]
359:
360:       #---------------#
361:       # calculando KS #
362:       #---------------#
363:       if(is.null(resumo_result)){
364:         ks_aux<- round(KS(prob,base_prob_perf[,.(IND_BOM)])
     ,4)
365:         nome_teste <- paste0("fitLD_mod",i,"_lambda_priori_it
     ",it)
366:         nova_linha <- cbind(nome_teste,ks_aux)
367:         resumo_result <- nova_linha
368:         colnames(resumo_result)<- c("teste","ks")
369:
370:       }else{
371:         ks_aux<- round(KS(prob,base_prob_perf[,.(IND_BOM)])
     ,4)
372:         nome_teste <- paste0("fitLD_mod",i,"_lambda_priori_it
     ",it)
373:         nova_linha <- cbind(nome_teste,ks_aux)
374:         colnames(nova_linha) <- colnames(resumo_result)
375:         resumo_result <- rbind(resumo_result,nova_linha)
376:       }
377:
378:
379:       #---------------------#
380:       # salvando resultados #
381:       #---------------------#
```

```
382:
383:         write.csv(base_prob_perf,
384:                   "C:/Users/Erick/Google Drive/USP/2020/
      Pesquisa/PowerLinks_LaplacesDemon/powerLinks_v3/
      base_prob_perfl_LD_POWERLINK.csv",
385:                   row.names = F)
386:
387:         write.csv(resumo_result,
388:                   "C:/Users/Erick/Google Drive/USP/2020/
      Pesquisa/PowerLinks_LaplacesDemon/powerLinks_v3/
      resumo_result_PL.csv",
389:                   row.names = F)
390:
391:
392:
393:         #---------------------#
394:         #   apagando aux      #
395:         #---------------------#
396:
397:         rm(prob,Fit,ks_aux,nome_teste,nova_linha)
398:
399:
400:         ind_tempo = ind_tempo+ aux_tempo
401:         print(paste0("concluido: ",100*round(ind_tempo,3),"%","
       completo"))
402:       }
403: }
404:
405: resumo_result = as.data.table(resumo_result)
406: resumo_result = resumo_result[order(-ks),]
407:
408: resumo_result
```


ICMC
USP
SÃO CARLOS