Universidade de Brasília - UnB
Instituto de Ciências Exatas - IE
Departamento de Estatística - EST

Trabalho de Conclusão de Curso 2

Análise e construção de variáveis para a melhoria de modelo de regressão logística de concessão de crédito do Banco de Brasília (BRB).

Ariston Dias de Farias

Orientador: Prof.º Alan Ricardo da Silva

Brasília

Setembro de 2022

Ariston Dias de Farias **17/0002039**

Análise e construção de variáveis para a melhoria de modelo de regressão logística de concessão de crédito do Banco de Brasília (BRB).

Relatório apresentado à disciplina Trabalho de Conclusão de Curso 2 do Bacharelado em Estatística, Departamento de Estatística, Instituto de Exatas, Universidade de Brasília, como parte dos requisitos necessários para o grau de Bacharel em Estatística.

Orientador: Prof. Dr. Alan Ricardo da Silva

Brasília

2022

Dedico este trabalho à minha família que me apoiou e esteve ao meu lado nos piores e melhores momentos. Em tempos tão difíceis ele me fizeram não desistir.

Agradecimentos

Agradeço a todos que passaram comigo esses momentos de esforço, dedicação e sofrimento. Em cada palavra há alguém me incentivando a escreve-la.

"*A fé é a certeza sem prova. A ciência é a probabilidade da certeza.*"

— Valter da Rosa

Resumo

Os modelos de concessão de crédito ou de risco de crédito são utilizados para tentar prever se o indivíduo tomador de crédito irá ou não honrar com suas dívidas. Durante anos o modelo mais utilizado no mercado foi o de regressão logística, tanto por sua facilidade quanto por sua explicabilidade. Com o aumento da potência dos computadores e o surgimento de novas técnicas, a regressão logística vem sendo substiuída largamente. O Banco de Brasília (BRB) é um exemplo de intituição financeira que utiliza o modelo de regressão logística para suas concessões. Esse documento realizou uma comparação entre os modelos de regressão logística simples, onde as variáveis do modelo foram incluídas diretamente e a regressão logística utilizando os componentes principais (ACP) como variáveis da regressão. A redução de dimensionalidade dos dados foi satisfatória mantendo uma porcentagem razoável da variância acumulada.

Com os resultado obtidos através do modelo de componentes principais verificou-se que o ele possui uma maior acurácia, um maior indicador K-S (Kolmogorov-Smirnov) e uma maior área da curva característica de operação (ROC), onde essa é a curva que ajuda a entender o desempenho do modelo. Esses são os principais indicadores utilizados nas análises de modelos de crédito. As variáveis escolhidas pelo modelo são apenas uma parte de todo o espectro de variáveis fazendo com que o modelo tenha larga possibilidade de melhoras e ganhos em seus indicadores.

Palavras-chaves: Regressão Logística. Amostragem. Modelo de Crédito. Componentes principais. Análise Multivariada. Banco de Brasília(BRB)

Abstract

Credit granting or credit risk models are used to try to predict whether or not the borrower will honor his debts. For years, the most used model in the market was the logistic regression model, both for its ease and for its explainability. With the increase in the power of computers and the emergence of new techniques, logistic regression has been largely replaced. Banco de Brasília (BRB) is an example of a financial institution that uses the logistic regression model for its concessions. This document made a comparison between logistic regression models and principal components analysis, using data from Banco de Brasília customers. The reduction in the dimensionality of the data was satisfactory, maintaining a reasonable percentage of the accumulated variance.

With the results obtained through the principal components model, it was verified that the model has a greater accuracy, a greater K-S indicator (Kolmogorov-Smirnov) and the greater area of the Receiver Operator Characteristc Curve (ROC) where this is the curve that helps understand the performance of the model. These are the main indicators used in the analysis of credit models. The variables chosen by the model are just a part of the entire spectrum of variables, making the model have a large possibility of improvements and gains in its indicators.

Palavras-chaves: Logistic Regression. Sampling. Credit Risk Model. Principal Components. Multvariate Analysis . Banco de Brasília (BRB) - Bank of Brasília

Lista de Tabelas

Lista de Figuras

Sumário

Capítulo 1

INTRODUÇÃO

Nos últimos anos o mercado financeiro tem ficado casa vez mais criterioso quanto a concessão e monitoramento do crédito (SERASA, 2021), fazendo-se necessário a implementação de novas estratégias e técnicas cada vez mais avançadas, visando sempre a melhoria e rapidez das concessões. Essas concessões são operações firmadas sob condições de incertezas e, devido a isso, sempre existirá uma probabilidade de perda dessas operações. Por mais que pareçam confiáveis, essa probabilidade de perda sempre existirá e se chama Risco de Crédito, conhecido também como *score* de crédito (Sicsú, 2009).

Com o cenário macro-econômico atual, completamente atípico e incerto devido a pandemia (SERASA, 2021), todos os dias novas empresas do mercado financeiro são criadas seguindo a onda digital, como explicitado pela Exame (2021) e com elas novas ideias, métodos mais avançados e um enorme crescimento de dados disponíveis vem tomando espaço no mercado. Diante desse cenário tão dinâmico e com a facilidade de se obter tantos dados, tem-se a necessidade de utilização de novas técnicas para avaliar e medir a quantidade de informação relevante que existe nesse grande volume de dados.

Os modelos de concessão de crédito atuais não utilizam apenas os dados fornecidos pelo usuário e seu comportamento na instituição, mas utilizam toda a gama de informações compartilhada no mercado financeiro, desde a porcentagem de utilização de limite de cartões de crédito à quantidade de protestos em cartórios nos últimos 6 meses (Jappelli e Pagano, 2000). Essas informações compartilhadas tornam as previsões dos modelos mais precisas, diminuem o risco de crédito e incentiva o tomador a não se endividar em várias instituições ao mesmo tempo (Jappelli e Pagano, 2000). Tullio Jappelli é atualmente o diretor e professor de finanças do Centro de Estudos em Economia e Finanças (CSEF) na Universidade de Nápoles Federico II e Marco Pagano é atualmente professor de finanças no Departamento de Econômia e Estatística na Universidade de Nápoles. No mercado existem grandes *bureaus* de crédito como SERASA, SPC, Boa Vista etc, e todos fazem parte da Agência Nacional de *Bureau* de Crédito (ANBC, 2021). Essa enorme quantidade de dados foi criada e disponibilizada ao longo dos anos para que as instituições financeiras criem modelos de concessão mais acertivos e com menor taxa de erros.

Com o amplo aumento da quantidade de dados, o mercado financeiro passou a utilizar em seus modelos de crédito esse grande volume de informações a fim de diminuir o risco do tomador de crédito. No entanto, algumas variáveis podem apresentar altas correlações entre elas, o que pode invalidar as suposições dos modelos de regressão logística. Devido a isso, a obtenção de informações mais relevantes se torna cada vez mais dificíl e custosa. Atualmente estão disponíveis diversas informações ligadas aos CPF dos clientes em diversas intituições.

Um outro modelo de crédito que o mercado passou a adotar em suas análises é o *Machine Learning*. Uma técnica que modifica seus parâmetros de construção de modelo com bases em seu fluxo de dados entrantes, reconhecendo padrões e a capacidade de aprender e fazer predições baseadas nesses dados.

Tendo em vista esse panorama, este trabalho buscará utilizar as técnicas de análise multivariada, para se extrair o máximo de informação, de diversas variáveis quantitativas simultaneamente, condensando e implementando com maior acertividade em modelos de regressão logística, já utilizados para a construção de modelos de concessão de crédito no BRB. Com modelos logísticos mais refinados, tem-se como consequência um menor índice de inadimplência, menores taxas de juros, produtos mais direcionados, menor exclusão de clientes que eram, anteriormente, classificados como não-pagadores ou que tenham baixo *score* e menor risco.

1.1 Objetivos

Esse trabalho tem como objetivo geral a construção ou identificação, através de técnicas de análise multivariada, de variáveis dentre as diversas disponíveis no mercado, para um refinamento do modelo de risco de crédito do BRB. As técnicas multivariadas consistem em análise que visam a interpretação, a organização, a sumarização e a extração de informações de bases de dados com inúmeras variáveis.

Os objetivos específicos são:

- Obter variáveis da SERASA, Sistema financeiro, etc. Informações não disponibilizadas pelos clientes do BRB;

- Criar novas componentes, baseadas nas variáveis obtidas de bases secundárias;

- Analisar a eficiência e o ganho efetivo da inclusão das novas variáveis criadas.

Capítulo 2

PRINCIPAIS TÉCNICAS DE ANÁLISE MULTIVARIADA

2.1 Introdução

Nesse Capítulo será feita uma revisão de algumas técnicas de análise multivariada que serão utilizadas no trabalho, a saber Análise Discriminate (Regressão Logística) e Análise de Componentes Principais. A Análise de Componentes Principais será utilizada para diminuir a quantidade de variáveis a ser analisada dentro de um modelo de Regressão Logística.

2.2 Análise Discriminante

A Análise Discriminante é uma técnica utilizada para discriminar dois ou mais grupos, utilizando funções lineares (Rencher, 2002). O objetivo principal dessa técnica é classificar observações em grupos que são conhecidos. Por exemplo, uma instituição financeira possui uma lista com correntistas classificadas como bom pagadores ou mau pagadores, e então deseja-se criar uma função, a partir de diversas variáveis, para discrimar novos tomadores de crédito. A Análise Discriminante avalia a probablidade de classificação em um ou outro grupo, cria métodos e ferramentas, para uma melhor

distinção entre esses grupos de populações e determina como classificar as novas observações nesses grupos criados.

Basicamente, para apenas dois grupos, tem-se as populações π_1 e π_2. As observações são classificadas, com base na probabilidade p e associadas as variáveis aleatórias $\mathbf{X} = [X_1, X_2, ..., X_p]$ sendo os valores da primeira classe de $\mathbf{X}$ para π_1 e para segunda classe de $\mathbf{X}$ para π_2 (Johnson e Wichern, 2002).

Em geral tem-se π_j populações com $j = 1, 2, \ldots, J$ e o objetivo é classificar o vetor de observações $\mathbf{X}$ em certos grupos, R_j, onde o erro de classificação errada é mínimo. O modelo discriminante é composto pela regra de separação das observações nas populações, tendo como objetivo principal minimizar o erro de classificações incorretas. A Figura 2.1 mostra um exemplo de dois grupos.

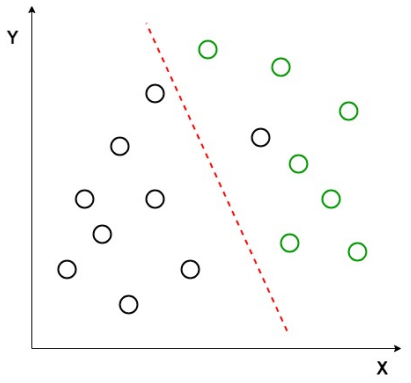

Figura 2.1: Exemplo de dois grupos

Na Figura 2.1 a reta é criada através das regras de separação da técnica de Análise Discriminante. Segundo Johnson e Wichern (2002), uma das formas mais recorrentes de criar uma função discriminante é por meio da Regressão Logística, que será visto a seguir.

2.2.1 Regressão Logística

Ao longo dos anos, os métodos de análise de Regressão Logística se tornaram parte integrante de qualquer análise de dados que se preocupa em descrever a relação entre uma variável resposta e uma, ou várias, variáveis explicativas. Em um modelo de Regressão Logística pode-se trabalhar com uma ou diversas variáveis explicativas. A Regressão Logística tem como objetivo principal identificar o melhor e mais parcimonioso ajuste entre as variáveis explicativas e a variável resposta (Hosmer e Lemeshow, 2000).

Em um modelo clássico de Regressão Logística, em geral, pode-se utilizar 2 categorias para a variável resposta Y, sendo eles 1 (sucesso) ou 0 (fracasso). Para o estudo de modelo de crédito em questão o valor 1 será usado para classificar o indivíduo inadimplentes (mau pagador) e o 0 para o indivíduo adimplente (bom pagador). Existem também os modelos logísticos multinomias, que consideram mais de uma categoria para a variável dependente Y, mas estes não serão tratados aqui. A probabilidade se ser um sucesso é denotada por $P(Y = 1)$ e a probabilidade de fracasso $P(Y = 0)$. Denota-se $P(Y = 1)$ sendo $\pi(x)$ para enfatizar que Y depende do valor de x dessa variável, sendo assim P(Y=1) = $\pi(x)$. Sendo as observações variáveis independentes com distribuição Bernoulli com parâmetro $\pi(x)$ (Agresti, 2015).

Para os modelos univariados, a equação linear assume a seguinte forma:

$$logit[\pi(x)] = log\left[\frac{\pi(x)}{1-\pi(x)}\right] = \alpha + \beta x \tag{2.1}$$

$$\pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} \tag{2.2}$$

A Equação (2.1) é chamada de função *Logito*.

Pode-se considerar que para cada unidade acrescida em x há um acréscimo de e^{β} em $\pi(x)$. Essa informação é de difícil intepretação, e para facilitar o entendimento foi criado o conceito de Chances ou *Odds* onde é a probabilidade de se obter sucesso dividida pela probabilidade de se obter fracasso, definida como:

$$\frac{\pi(x)}{1-\pi(x)} = exp(\alpha + \beta x) = e^{\alpha}(e^{\beta})^{x} \tag{2.3}$$

Quando a *Odds* é maior que 1 a Chance de sucesso é maior que a Chance de fracasso. Quando a *Odds* é menor que 1 a Chance de sucesso é menor que a Chance de fracasso. Com o conceito de Chance criado, há a necessidade de se medir a associação entre as Chances de sucesso e as Chances de fracasso de duas variáveis com isso é introduzido o termo de Razão de Chances ou *Odds Ratio*(OR). Sendo a probabilidade de sucesso igual a π_i e a probabilidade de fracasso 1 - π_i a Razão de Chances nada mais é que a divisão entre a Chance de sucesso e a Chance de fracasso das variáveis escolhidas para a comparação(Agresti, 2015). A OR é definida como:

$$OR = \frac{\pi_1/[1-\pi_1]}{\pi_2/[1-\pi_2]} = \theta, \quad i = 1, 2 \tag{2.4}$$

Quando as variáveis X_1 e X_2 são independentes $\pi_1 = \pi_2$, então $\text{Chance}_1 = \text{Chance}_2$ e $\theta = 1$. No caso once $\theta > 1$ então a $\text{Chance}_1 > \text{Chance}_2$.

Com o conceito do modelo univariado e bivariado há a necessidade de se expandir os conceitos para o modelo multivariado. O modelo multivariado ocorre quando existem duas ou mais variáveis explicativas relacionadas com a variável resposta.

Segundo Johnson e Wichern (2002) tendo p variáveis independentes compondo o vetor $\mathbf{X'} = (x_1, x_2, x_3, ..., x_p)$ e assumindo que a probabilidade $P(Y = 1|\mathbf{x}) = \pi(x)$,

seja a probabilidade condicional, onde Y está condicionado a x. O modelo de regressão logística múltipla será da forma:

$$\mathbf{Y} = g(\mathbf{X}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_n \tag{2.5}$$

Sendo escrito na forma matricial como:

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}; \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}; \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

$$\underbrace{\mathbf{Y}}_{n\times 1} = \underbrace{\boldsymbol{\beta}}_{p\times 1}\ \underbrace{\mathbf{X}}_{n\times p}$$

Sendo $\mathbf{Y}$ o vetor de variáveis resposta, $\boldsymbol{\beta}$ o vetor de parâmetros e $\mathbf{X}$ a matriz de valores.

No caso do modelo múltiplo, a função *logito* será:

$$logit[\pi(x)] = log\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p = g(\mathbf{x}) \tag{2.6}$$

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_p x_p}} = \frac{e^{g(\mathbf{x})}}{1+e^{g(\mathbf{x})}} \tag{2.7}$$

A Equação (2.7) é a função logito múltipla.

Tendo os parâmetros $\beta_1, \beta_2, \ldots, \beta_n$ e sendo β_1 que representa a mudança em $\pi(x)$ para cada unidade acrescida em x_1, β_2 a mudança em x_2 e assim por diante. Para modelos logito tem-se apenas 2 valores para a variável Y, 0 = fracasso e 1 = sucesso como mostrado na Figura a seguir.

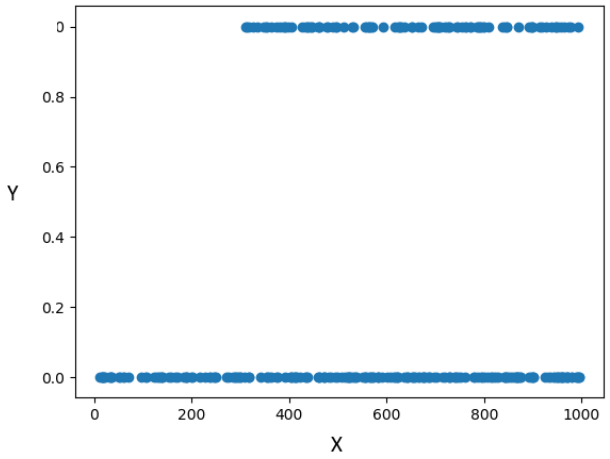

Figura 2.2: Exemplo de um gráfico de dispersão entre X e Y

A Figura 2.2 mostra o gráfico de dispersão dos pontos onde o eixo X são as variáveis independentes e o eixo Y a variável resposta.

Para a construção da curva na Figura 2.3, usa-se os valores das probabilidades $\pi(x)$ onde a inclinação da reta que tangencia o ponto x na curva é igual a $\beta\pi(x)$[1-$\pi(x)$].

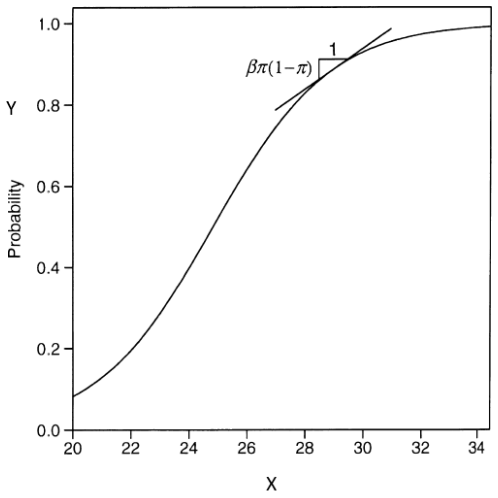

Figura 2.3: Aproximação linear da curva de regressão logística

Para o ponto em destaque, a Figura 2.3 mostra a taxa de mudança naquele ponto. No ponto onde $\pi(x) = 0{,}50$ tem-se uma inclinação de $\beta(0,5)(0,5) = 0,25\beta$ que é o

Nível efetivo médio (Agresti, 2015).

Método de estimação

Nos modelos de Regressão Logística a inferência estatística ajuda a medir a significância e os efeitos das variáveis explicativas sobre a variável reposta. As estatísticas de Wald e a Razão de Máxima Verossimilhança são amplamente utilizadas devido a sua significância e relevância nos modelos (Agresti, 2015).

Máxima Verossimilhança

A título de simplificação a Máxima Verossimilhança será mostrada no exemplo univariado, mas será utillizada o modelo múltiplo. Segundo Casella e Berger (2018), na função de distribuição acumulada (fda) de uma distribuição logística $F(w) = e^w/(1 + e^w)$, a função $\pi(x)$ assume a forma $\pi(x) = F(\beta_0 + \beta_1 x)$, assim pode-se considerar que $F_i(x) = F(\beta_0 + \beta_1 x)$, então a função de Verossimilhança é igual a:

$$L(\alpha, \beta|y) = \prod_{i=1}^{n} \pi(x_i)^{y_i}(1 - \pi(x_i))^{1-y_i} = \prod_{i=1}^{n} F_i^{y_i}(1 - F_i)^{1-y_i} \tag{2.8}$$

Toma-se o log da Verossimilhança pois há uma maior facilidade algébrica sem perda das propriedades da equação. O log da Máxima Verossimilhança é dado por:

$$log(L(\alpha, \beta|y)) = \sum_{i=1}^{n} \left\{ log(1 - F_i) + y_i log\left(\frac{F_i}{1 - F_i}\right) \right\} \tag{2.9}$$

Para obter as equações de Verossimilhança toma-se a derivada do log da Verossimilhança em relação a α e β. Lembrando que a derivada de $F(w)/d(w)$, é a função de densidade de probabilidade, ou seja, $F(w)/d(w) = f(w)$. Sendo a derivada do *log*

da Verossimilhança em relação a α:

$$\frac{\partial}{\partial\alpha}log(L(\alpha,\beta|y)) = \sum_{i=1}^{n}(y_i - F_i)\frac{f_i}{F_i(1-F_i)} \tag{2.10}$$

De maneira similiar a derivada em relação a β, tem-se:

$$\frac{\partial}{\partial\beta}log(L(\alpha,\beta|y)) = \sum_{i=1}^{n}(y_i - F_i)\frac{f_i}{F_i(1-F_i)}x_i \tag{2.11}$$

Os estimadores de Máxima Verossimilhança são gerados resolvendo as equações diferenciais e igualando a zero, porém são equações não lineares em α e β e devem ser resolvidas numericamente (Casella e Berger, 2018).

Os estimadores de Máxima Verossinilhança podem ser usados para a realização do teste do parâmetro β nos modelos de regressão logística. Para Hosmer e Lemeshow (2000) tem-se no modelo a hipótese de $H_0 : \beta = 0$ contra $H_1 : \beta \neq 0$. O EMV do modelo simples, onde $\beta = 0$, é igual a l_0. O EMV do modelo completo, com todos os parâmetros, é igual a l_1, então a Razão de Máxima Verossimilhança é dada por:

$$2log(l_1/l_0) \tag{2.12}$$

Para valores grandes dessa estatística e $p-valores$ pequenos rejeita-se H_0 onde há evidências que o parâmetro β é diferente de 0 como suposto pela hipótese alternativa, H_1 (Hosmer e Lemeshow, 2000).

Teste de Wald

Nos modelos lineares o parâmetro β possue efeito e relevância sobre a variável explicativa, indicando peso e o sentido do gráfico, crescente ou descrescente. Considerando que β possa ser 0, sem relevância, tem-se a hipótese nula de $H_0 : \beta = 0$ contra a hipótese alternativa $H_1 : \beta \neq 0$. Sendo $\hat{\beta}$ o estimador de Máxima Verossimilhança

(EMV) para a distribuição e $EP(\hat{\beta})$ o erro padrão de $\hat{\beta}$. Sob H_0 tem-se a estatística do teste (Hosmer e Lemeshow, 2000). A estatística de Wald é dada por:

$$z = \frac{\hat{\beta} - \beta_0}{EP(\hat{\beta})} \tag{2.13}$$

Essa estatística possui, aproximadamente, uma distribuição normal padrão, o que equivale dizer que z^2 possui distribuição χ^2 com 1 grau de liberdade e se chama Estatística de Wald.

Um conceito de grande importância a ser introduzido é o conceito de *Deviance*. Dentre todos os modelos canditatos, o mais complexo possível será o modelo onde todos os coeficientes β são escolhidos. Agora considere o modelo de interesse onde há apenas alguns parâmetros escolhidos, não todos. Segundo Agresti (2015) para o modelo saturado, denota-se o o log da Máxima Verossimilhança de L_S e para o modelo escolhido de L_M. Sabendo que o log da Máxima Verossimilhança, do modelo saturado, sempre será maior do que qualquer outro modelo com menos parâmetros, ou mais simples, a *Deviance* é igual a:

$$Deviance = 2(L_S - L_M) = G^2 \tag{2.14}$$

A Equação (2.14) nada mais é que a razão do log da Máxima Verossimilhança entre o modelo saturado e o modelo candidato, indicando a distância entre o modelo saturado e o modelo candidato. Essa estatística possue distribuição Chi quadrado (χ^2) com com $r - s$ graus de liberdade, onde r é a quantidade de graus de liberdade do modelo reduzido ou candidato e s é a quantidade de graus de liberdade do modelo saturado ou modelo cheio (Hardle e Simar, 2015). Assim a distribuição é calculada

da seguinte forma:

$$G^2 \sim \chi^2_{r-s} \tag{2.15}$$

Tem-se as seguintes hipóteses para o teste:

H_0 : Modelo reduzido, com r graus de liberdade.

H_1 : Modelo saturado, com s graus de liberdade.

Sob H_0:

$$G^2_{H_0} - G^2_{H_1} \sim \chi^2_{r-f} \tag{2.16}$$

A rejeição de H_0 acontece quando o modelo candidato, com menos parâmetros, não é escolhido. Sendo um α escolhido previamente, quando o $p - valor$ é menor que o α escolhido (Hardle e Simar, 2015). A probabilidade do teste é dada da seguinte forma:

$$P\{\chi^2_{r-f} > (G^2_{H_0} - G^2_{H_1})\} \tag{2.17}$$

Ajuste e teste de significância do modelo

Considerando a situação inicial onde deseja-se separar a população $\mathbf{X}$ em dois grupos distintos, π_1 e π_2 e classificar os novos dados em um desses grupos. Os valores dessa população são separados ou classificados com, base em suas medições, por exemplo, em p variáveis aleatórias associadas. Os valores observados de $\mathbf{x}$ diferem até certo ponto entre π_1 e π_2. Pode-se pensar nos valores da primeira classe como sendo a população de valores $\mathbf{x}$ de π_1 e os da segunda classe como a população de valores $\mathbf{x}$ de π_2. Essas populações podem ser descritas através de suas densidades de probabilidade $f_1(\mathbf{x})$ e $f_2(\mathbf{x})$ (Johnson e Wichern, 2002).

Segundo (Johnson e Wichern, 2002) para um perfeito ajuste do modelo de Regressão Logística a regra discriminante deve poder classificar com os menores erros possíveis. No modelo de Regressão Logística há dois tipo de erros, chamados erros de classificação, eles se apresentam de duas maneiras:

- Erro I = Quando o modelo classifica na região R_2, porém ele pertence a região R_1
- Erro II = Quando o modelo classifica na região R_1, porém ele pertence a região R_2.

Para se reescrever os erros citados em função de suas probabilidades, define-se R_1 como sendo a região onde classifica-se as observações como π_1 e R_2 em π_2 (Johnson e Wichern, 2002). Para a probabilidade do Erro I tem-se a seguinte probabilidade condicional:

$$P(2|1) = P(\mathbf{X} \in R_2|\pi_1) = \int_{R_2} f_1(\mathbf{x})d\mathbf{x} \tag{2.18}$$

Essa é a probabilidade de se classificar em π_2 quando a observação pertence à π_1. O mesmo vale para $P(1|2)$ sendo a probabilidade de se classificar em π_1 quando a observação pertence à π_2, sendo a probabilidade do Erro II igual a:

$$P(1|2) = P(\mathbf{X} \in R_1|\pi_2) = \int_{R_1} f_2(\mathbf{x})d\mathbf{x} \tag{2.19}$$

Quando há classificações em duas regiões as probabilidades podem ser classificadas da seguinte maneira:

Probabilidade de se classificar corretamente em $\pi_1 =$

$$P(\mathbf{X} \in R_1|\pi_1)P(\pi_1) = P(1|1)p_1 \tag{2.20}$$

Probabilidade de se classificar corretamente em $\pi_2 =$

$$P(\mathbf{X} \in R_2|\pi_2)P(\pi_2) = P(2|2)p_2 \tag{2.21}$$

Probabilidade de se classificar em π_1 sendo que petence a $\pi_2 =$

$$P(\mathbf{X} \in R_1|\pi_2)P(\pi_2) = P(1|2)p_2 \tag{2.22}$$

Probabilidade de se classificar em π_2 sendo que petence a $\pi_1 =$

$$P(\mathbf{X} \in R_2|\pi_1)P(\pi_1) = P(2|1)p_1 \tag{2.23}$$

$P(1|1)p_1$ = Probabilidade de se classificar um indivíduo inadimplente dado que ele é inadimplente.

$P(2|2)p_2$ = Probabilidade de se classificar um indivíduo adimplente dado que ele é adimplente.

$P(1|2)p_2$ = Probabilidade de se classificar um indivíduo inadimplente dado que ele é adimplente.

$P(2|1)p_1$ = Probabilidade de se classificar um indivíduo adimplente dado que ele é inadimplente.

São as equações de probabilidades possíveis para duas regiões possíveis de classificação (Johnson e Wichern, 2002). Essas classificações podem ser exemplificadas em uma matriz, chamada matriz de confusão.

		Região Classificada		
		π_1	π_2	
Região verdadeira	π_1	n_{1C}	n_{1E}	n_1
	π_2	n_{2E}	n_{2C}	n_2

onde:

n_{1C} = Número de observações de π_1, classificados corretamente em π_1;

n_{1E} = Número de observações de π_1, classificados erroneamente em π_2;

n_{2C} = Número de observações de π_2, classificados corretamente em π_2;

n_{2E} = Número de observações de π_2, classificados erroneamente em π_1.

Essa matriz é amplamente utilizada para calcular importantes indicadores de ajuste do modelo que são Especificidade e Sensitividade (Agresti, 2015). A Especificidade e Sensitividade são dadas como:

$$Especificidade = \frac{n_{2C}}{n_{2E} + n_{2C}} = \frac{n_{2C}}{n_2}; \tag{2.24}$$

$$Sensitividade = \frac{n_{1C}}{n_{1C} + n_{1E}} = \frac{n_{1C}}{n_1} \tag{2.25}$$

A Especificidade é a probabilidade de se classificar corretamente uma observação de π_2 em π_2 e a Sensitividade é a probabilidade de se classificar corretamente uma obervação de π_1 em π_1 (Agresti, 2015).

2.3 Análise de Componentes Principais

Segundo Johnson e Wichern (2002) a Análise de Componentes Principais (ACP) é uma técnica multivariada que visa explicar a matriz de variância-covariância de um conjunto de dados quantitativos através de algumas combinações lineares de suas variáveis, e tem como objetivo a redução da dimensionalidade, facilitar interpretações e reduzir a multicolinearidade.

A ideia principal da ACP é criar uma transformação linear de um conjunto de variáveis quantitativas, inicialmente correlacionadas, em um conjunto com menos variáveis e não correlacionadas entre si. Supondo que existam p variáveis, onde dimensionalmente e geometricamente é difícil de se obter qualquer informação, a técnica de componentes principais realiza uma redução para k componentes ($k < p$), onde essa quantidade possue a maior quantidade de informação relevante (Johnson e Wichern, 2002).

Uma das formas de se medir a quantidade de informação é através da variância, que para ACP, é a quantidade de informação em uma combinação linear. A Análise de Componentes Principais visa a maximização dessa variância em uma combinação linear. Sob o conceito geométrico essa combinação linear representa a seleção de um novo sistema de coordenadas, obtido através de rotações específicas do sistema original com X_1, X_2, ..., X_p onde os componentes representam os vetores com direção da máxima variância (Johnson e Wichern, 2002).

De acordo com Johnson e Wichern (2002) essa técnica pode ser utilizada para realizar agrupamentos de indivíduos segundo sua variância, ou seja, ela analisa o comportamento da população segundo a variação de suas característsticas. De fato as variáveis da matriz $\mathbf{X}$, possuem variâncias, covariâncias e correlações entre si, tendo em mente essa característica, cria-se a matriz de variância-covariância dessas variáveis, denotada por:

$$\sum = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \dots & \sigma_{2p} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \dots & \sigma_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \sigma_{n3} & \dots & \sigma_{np} \end{bmatrix} \tag{2.26}$$

Para as p-variáveis há uma matriz de variância-covariância ($\sum$), uma matriz de correlação($\boldsymbol{\rho}$) e um vetor média expresso por $\boldsymbol{\mu} = [\mu_1, \mu_2, \ldots, \mu_p]^T$. Associados a essas matrizes tem-se os pares de autovalores-autovetores (λ_1, e_1), (λ_2, e_2), $\ldots$, (λ_p, e_p) onde $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$ (Johnson e Wichern, 2002).

Os auto-valores e auto-vetores são usados para realizar a decomposição espectral onde essa decomposição é uma maneira de se reescrever uma matriz em função de seus auto-valores e auto-vetores. Tendo a matriz $\mathbf{A}$ como exemplo, pode-se reescreve-la como $\mathbf{A} = \sum_{i=1}^{n} \lambda_i e_i e_i'$ em função de seus auto-valores (λ_i) e auto-vetores (e_i). Para a obtenção de λ_i deve-se calcular o determinante da equação definida por (Johnson e Wichern, 2002):

$$det[\mathbf{A} - \boldsymbol{\lambda}\mathbf{I}] = 0 \tag{2.27}$$

onde $\mathbf{I}$ é a matriz identidade de $\mathbf{A}$.

Para obtenção dos auto-vetores de $\mathbf{A}(\mathbf{e}_i)$ deve-se obter as raízes dos sistema de equações gerados por:

$$\mathbf{A}\mathbf{x} = \boldsymbol{\lambda}\mathbf{x} \tag{2.28}$$

onde $\mathbf{x}$ é o vetor característico da matriz $\mathbf{A}$ associado ao auto-valor λ. Os auto-vetores serão:

$$\mathbf{e} = \frac{\mathbf{x}}{\sqrt{\mathbf{x}'\mathbf{x}}} \tag{2.29}$$

Para exemplificar as equações citadas considere a matriz $\mathbf{A}$ sendo:

$$\mathbf{A} = \begin{bmatrix} 2.2 & 0.4 \\ 0.4 & 2.8 \end{bmatrix} \tag{2.30}$$

Usando a Equação (2.27), tem-se:

$$det[\mathbf{A} - \lambda\mathbf{I}] = \lambda^2 - 5\lambda + 6.16 - 0.16 = (\lambda - 3)(\lambda - 2) \tag{2.31}$$

onde obtem-se os auto-valores $\lambda_1 = 3$ e $\lambda_2 = 2$.

Para os auto-vetores tem-se $\mathbf{e}_1' = [1/\sqrt{5}, 2/\sqrt{5}]$ e $\mathbf{e}_2' = [2/\sqrt{5}, -1/\sqrt{5}]$, consequentemente pode-se reescrever a matriz $\mathbf{A}$ como:

$$\mathbf{A} = \begin{bmatrix} 2.2 & 0.4 \\ 0.4 & 2.8 \end{bmatrix} = 3 \begin{bmatrix} \frac{1}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \end{bmatrix} + 2 \begin{bmatrix} \frac{2}{\sqrt{5}} \\ \frac{-1}{\sqrt{5}} \end{bmatrix} \begin{bmatrix} \frac{2}{\sqrt{5}} & \frac{-1}{\sqrt{5}} \end{bmatrix} \tag{2.32}$$

$$= \begin{bmatrix} 0.6 & 1.2 \\ 1.2 & 2.4 \end{bmatrix} + \begin{bmatrix} 1.6 & -0.8 \\ -0.8 & 0.4 \end{bmatrix} = \begin{bmatrix} 2.2 & 0.4 \\ 0.4 & 2.8 \end{bmatrix} \tag{2.33}$$

Nota-se claramente o resultado da decomposição espectral, onde $\mathbf{A}$ pode ser escrita em função de seus auto-valores e auto-vetores (Johnson e Wichern, 2002).

Um dos objetivos da ACP é obter componentes que não são correlacionados entre si, ou seja, que as combinações lineares de Y_1, Y_2, $\ldots$, Y_p não sejam correlacionadas e que tenham a maior variância possível (Johnson e Wichern, 2002). Para isso deve-se identificar combinações lineares de Y_1, Y_2, $\ldots$, Y_p que não sejam correlacionadas entre si e que tenham a maior variância possível, isso significa que as combinações lineares devem ser do tipo:

$$\begin{aligned} Y_1 &= \mathbf{a}_1'\mathbf{X} = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p \\ Y_2 &= \mathbf{a}_2'\mathbf{X} = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p \\ &\vdots \qquad\qquad\qquad \vdots \\ Y_p &= \mathbf{a}_p'\mathbf{X} = a_{p1}X_1 + a_{p2}X_2 + \ldots + a_{pp}X_p \end{aligned}$$

Assim a variância e a covariância são escritas da seguinte forma:

$$Var(Y_i) = Var(\mathbf{a}_i'\mathbf{X}) = \quad \mathbf{a}_i' \sum \mathbf{a}_i, \quad i = \{1,\, 2,\, \ldots,\, p\} \tag{2.34}$$

$$Cov(Y_i,\, Y_k) = Cov(\mathbf{a}_i'\mathbf{X},\, \mathbf{a}_k'\mathbf{X}) = \quad \mathbf{a}_i' \sum \mathbf{a}_k, \quad i,\, k = \{1,\, 2,\, \ldots,\, p\} \tag{2.35}$$

Os componentes principais são combinações lineares de $Y_1,\, Y_2,\, \ldots,\, Y_p$ onde não há correlações entre elas e suas variâncias são as máximas possíveis. Para a criação dessas combinações lineares com as características citadas serão usados os conceitos de auto-valores e auto-vetores citados (Johnson e Wichern, 2002). Dessa forma para a matriz de covariância $\sum$ e o vetor $\mathbf{X}' = [x_1,\, x_2,\, x_3,\, \ldots,\, x_p]$ associado, tem-se o i-ésimo par de auto-valor e auto-vetor da matriz de covariância $\sum$ igual a $(\lambda_i,\, \mathbf{e}_i)$, onde $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$. Então os componentes principais serão expressos na forma:

$$\begin{aligned}
Y_1 &= e_{11}x_1 + e_{12}x_2 + e_{13}x_3 + \ldots + e_{1p}x_p = \mathbf{a}_1'\mathbf{X} \\
Y_2 &= e_{21}x_1 + e_{22}x_2 + e_{23}x_3 + \ldots + e_{2p}x_p = \mathbf{a}_2'\mathbf{X} \\
&\vdots \qquad \vdots \\
Y_i &= e_{i1}x_1 + e_{i2}x_2 + e_{i3}x_3 + \ldots + e_{ip}x_i = \mathbf{a}_i'\mathbf{X} \\
Y_i &= \mathbf{e}_i'\mathbf{X}, \quad i = 1, 2, \ldots, p
\end{aligned}$$

Com $\mathbf{Y}$ podendo ser escrito na forma matricial:

$$\mathbf{Y} = \mathbf{e}\mathbf{X}', \quad \text{para } \mathbf{e} = \begin{bmatrix} e_1' \\ \vdots \\ e_p' \end{bmatrix} \tag{2.36}$$

Sendo as combinações lineares escolhidas com as formas exemplificadas, as variâncias e covariâncias dos componentes são:

$$Var(Y_i) = \quad \mathbf{e}_i' \sum \mathbf{e}_i = \quad \lambda_i \tag{2.37}$$

$$Cov(Y_i,\, Z_k) = \quad \mathbf{e}_i' \sum \mathbf{e}_k = \quad \mathbf{e}_i'\lambda_k\mathbf{e}_k = \lambda_k\mathbf{e}_i'\mathbf{e}_k = 0 \tag{2.38}$$

Na Equação (2.37), λ_i é a variância máxima para a componente Y_i.

Com a definição dos pares de auto-valores e auto-vetores uma das conclusões mais importantes, para análise de componentes principais, é que a soma dos auto-valores é igual a soma das variâncias das variáveis, ou seja, a soma da diagonal da matriz $\sum$ é igual a soma dos auto-valores, apresentada da seguinte forma (Johnson e Wichern, 2002):

$$\sigma_{11} + \sigma_{22} + \sigma_{33} + \ldots + \sigma_{pp} = \sum_{i=1}^{p}(Var(X_i)) = \lambda_1 + \lambda_2 + \lambda_3 + \ldots + \lambda_p = \sum_{i=1}^{p}(\lambda_i)$$

Consequentemente pode-se afirmar que a proporção da k-ésima variância é a mesma do k-ésimo auto-valor, sendo assim:

$$\frac{\sigma_k}{\sigma_{11} + \sigma_{22} + \sigma_{33} + \ldots + \sigma_{pp}} = \frac{\lambda_k}{\lambda_1 + \lambda_2 + \lambda_3 + \ldots + \lambda_p} \tag{2.39}$$

A proporção da quantidade de informação contida na $k-$ésima variância é a mesma proporção do $k-$ésimo auto-valor (Johnson e Wichern, 2002).

Um dos problemas enfrentados na Análise de Componentes Principais, através de seus auto-valores e auto-vetores, são as escalas das variáveis da matriz $\mathbf{X}$, onde algumas podem variar de 0 a 1, outras de -100 a 100 ou de 0 a 1000, sendo assim, totalmente desproporcionais. Isso faz com que as variâncias da matriz $\sum$ sejam muito diferentes inviabilizando a criação de bons componentes principais, pois algumas variáveis vão contribuir muito menos que outras. Quando há uma mudança nas escalas de um ou mais valores de $\mathbf{X}$, a forma da distribuição dos pontos irá mudar, o que implica em encontrar novos componentes principais (Hardle e Simar, 2015).

Sabendo que os componentes não são invariantes em escala há uma preocupação com as unidades que as variáveis estão expressas, quando estão expressas na mesma

unidade de medida tudo ocorre bem, porém, na maioria dos casos, isso não ocorre. Uma das soluções adotadas é a padronização das variáveis antes de se extrair os auto-valores e auto-vetores, sendo a normalização obtida através das transformações:

$$Z_1 = \frac{(X_1-\mu_1)}{\sqrt{\sigma_{11}}} \tag{2.40}$$

$$Z_2 = \frac{(X_1-\mu_2)}{\sqrt{\sigma_{22}}} \tag{2.41}$$

$$\vdots \quad \vdots$$

$$Z_p = \frac{(X_p-\mu_p)}{\sqrt{\sigma_{pp}}} \tag{2.42}$$

A Equação (2.40) implica, segundo (Johnson e Wichern, 2002), que ao invés de se obter os componentes da matriz de variância-covariância $\sum$, encontra-se as componentes da matriz de correlação, $\boldsymbol{\rho}$, descrita como:

$$\boldsymbol{\rho} = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \dots & \rho_{1p} \\ \rho_{21} & 1 & \rho_{23} & \dots & \rho_{2p} \\ \rho_{31} & \rho_{32} & 1 & \dots & \rho_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \rho_{p3} & \dots & 1 \end{bmatrix} \tag{2.43}$$

Para a matriz de correlação a proporção de auto-valores não segue a mesma regra da matriz de variância-covariância. A proporção segue a seguinte regra:

$$\text{(Proporção da população normalizada)} = \frac{\lambda_k}{p}, \text{para } k = \{1, 2, 3, \dots, p\} \tag{2.44}$$

Segundo Johnson e Wichern (2002) a matriz de variância-covariância produz uma proporção entre os auto-valores diferente da proporção da matriz de correlação.

Johnson e Wichern (2002) apresenta um exemplo que mostra a divergência entre as proporções. Esse exemplo é feito da seguinte forma. Sendo a matriz de variância-covariância igual a:

$$\sum = \begin{bmatrix} 1 & 4 \\ 4 & 100 \end{bmatrix}$$

E a matriz de correlação igual a:

$$\boldsymbol{\rho} = \begin{bmatrix} 1 & 0.4 \\ 0.4 & 1 \end{bmatrix}$$

Os pares auto-valores e auto-vetores da matriz de variância-covariância $\sum$ são:

$$\lambda_1 = 100.16, \quad \mathbf{e}_1' = [0.040, 0.999] \tag{2.45}$$

$$\lambda_2 = 0.84, \quad \mathbf{e}_2' = [0.999, -0.040] \tag{2.46}$$

Os pares auto-valores e auto-vetores da matriz de correlação $\boldsymbol{\rho}$ são:

$$\lambda_1 = 1.4, \quad \mathbf{e}_1' = [0.707, 0.707] \tag{2.47}$$

$$\lambda_2 = 0.6, \quad \mathbf{e}_2' = [0.707, -0.707] \tag{2.48}$$

Os componentes principais extraídos da matriz $\sum$ são:

$$Y_1 = 0.040X_1 + 0.999X_2 \tag{2.49}$$

$$Y_2 = 0.999X_1 + 0.040X_2 \tag{2.50}$$

Os componentes principais extraídos da matriz $\boldsymbol{\rho}$ são:

$$Y_1 = 0.707Z_1 + 0.707Z_2 = 0.707(X_1 - \mu_1) + 0.707(X_2 - \mu_2) \tag{2.51}$$

$$Y_2 = 0.707Z_1 - 0.707Z_2 = 0.707(X_1 - \mu_1) + 0.707(X_2 - \mu_2) \tag{2.52}$$

Para as proporções dos auto-valores na matriz $\sum$, tem-se:

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{100.16}{101} = 0.992 \tag{2.53}$$

Nota-se que λ_1 é muito maior que λ_2 fazendo com que o primeiro componente tenha uma dominância sobre o segundo. Isso implica que o segundo componentes

não expressa bem parte da variância da matriz $\sum$. Há uma desproporcionalidade entre λ_1 e λ_2 para a matriz de variância-covariância $\sum$.

Para os componentes principais extraídos da matriz $\boldsymbol{\rho}$ com dados normalizados, tem-se:

$$\rho_{Y_1,Z_1} = e_{11}\sqrt{\lambda_1} = 0.707\sqrt{1.4} = 0.837 \tag{2.54}$$

$$\rho_{Y_1,Z_1} = e_{21}\sqrt{\lambda_1} = 0.707\sqrt{1.4} = 0.837 \tag{2.55}$$

As variáveis da matriz $\boldsymbol{\rho}$, para um sistema normalizado, contribuem igualmente, diferentemente das contribuições das variáveis de $\sum$.

Para a matriz de correlação $\boldsymbol{\rho}$ tem-se a seguinte proporção:

$$\frac{\lambda_1}{p} = \frac{1.4}{2} = 0.7 \tag{2.56}$$

Na matriz de correlação $\boldsymbol{\rho}$ nota-se um melhor equilíbrio nas proporcionalidades dos auto-valores e uma distribuição mais igualitária das variáveis normalizadas de cada componente principal. Esse fator evidencia que para dados desbalanceados, ou em escalas diferentes, o uso dos dados normalizados é melhor do que o uso dos dados em sua forma natural e desbalanceada (Johnson e Wichern, 2002).

Capítulo 3

MATERIAIS E MÉTODOS

3.1 Introdução

Este Capítulo tem a finalidade de mostrar o material utilizado e a metodologia aplicada para alcançar o objetivo proposto de redução da quantidade de variáveis, criação de componentes e melhoria do modelo logístico de concessão de crédito. Para isso serão coletadas diversas variáveis externas do mercado financeiro, para se aplicar a Analise de Componentes Principais realizando a redução em sua quantidade e minimizando a colinearidade entre elas.

3.2 Materiais

Os dados foram obtidos com a ajuda da SERASA. Esses dados compõe sua base de dados principal, disponibilizada ao mercado. Uma lista com uma amostra de clientes do BRB foi criada e enviada à SERASA, que por sua vez agregou todas as suas variáveis e enviou de volta com todas essas informações.

No modelo logístico já utilizado pelo BRB tem-se as seguintes variáveis:

- Valor do *Score* da Serasa;

- Quantidade Operações Vencidas (SFN/SCR-BACEN);
- Valor Total de Operações na Modalidade 19 (SFN/SCR-BACEN);
- Saldo Médio Mensal em Poupança;
- Código do Segmento do Cliente;
- Grau de Instrução;
- Saldo Médio Mensal em CDB;
- Dias de atraso nos últimos 30 dias na Cartão;
- Dias de atraso nos últimos 180 dias no Conglomerado BRB;
- Média do Valor Mensal de Movimentação Financeira (R$);
- Tempo de Abertura de Conta, em Meses;
- Tipo de Servidor.

3.3 Métodos

Com a obtenção dos dados da SERASA, primeiramente será feita uma matriz de correlação entre todas as variáveis afim de indentificar os problema de multicolinearidade para que então possa ser feita a ACP. A ACP será feita de duas maneiras:

- Utilizando todas as variáveis, independente de correlações existentes.
- Utilizando apenas as variáveis correlacionadas.

Para a criação das componentes principais foram utilizadas 60 variáveis que estão descritas no apêndice

Após a criação das componentes principais, serão incluídas no modelo logístico, para que os ajustes sejam feitos. Serão feitos dois ajustes, um para cada método utilizado, para que por fim sejam comparados com o modelo vigente e analisando se houve, ou não, melhorias significativas que justifiquem a criação e implementação das componentes. Para a comparação com a regressão logística vigente as variáveis utilizadas na regressão serão incluidas na criação dos componentes principais.

3.4 Extração

Os dados extraídos das bases do BRB são informações creditícias tanto do próprio BRB quanto da SERASA. A safra utilizada foi a de novembro de 2021. O evento *default* foi definido como uma visão de 6 meses após a data de referência para que possam ser analisados até maio de 2022. Uma segunda definição de *default* é a contratação dos produtos de renegociação, um indicador da deterioração da operação contratada, também analisada nesse período. Caso o cliente possua qualquer atraso maior que 60 dias, ou contratação de renegociação (RENEG), ambos no periodo de 6 meses, ele será classificado como evento *default*.

3.5 Limpeza

Foi feita uma análise de variância em todas as variáveis selecionadas e retiradas todas com variabilidade iguais a zero, onde tinha-se apenas um valor em todos os registros. Devido ao grande volume pode-se excluir todos os registros que não

possuiam valores válidos paras as variáveis escolhidas.

3.6 Manipulação

A parte prática desse trabalho foi feita nos servidores do BRB, extraindo-se apenas os resultados cumprido-se todas as leis da LGPD. A base total continha 1.119.789 registros onde 20%(223.957) foi separada para testar o modelo e 80%(895.832) foi separada para a criação do modelo.

Capítulo 4

ANÁLISE DOS RESULTADOS

4.1 Introdução

Este Capítulo tem como objetivo mostrar os tratamento e análise dos dados extraídos. O evento *default* foi definido como atraso maior que 60 dias em um período de 6 meses para cada operação de crédito de um determinado cliente ou a contratação de alguma operação de renegociação (RENEG). Para toda a análise dos dados foram utilizados 1.119.789 registros extraída aleatoriamente da base completa.

4.2 Resultados

Primeiramente foi construída a matriz de correlação para a obtenção dos auto-valores e auto-vetores. Através dos auto-valores foram calculadas as variâncias compartilhadas de cada componentes. Com as variâncias calculadas foi feita a soma acumulada de cada uma para se obter a quantidade de informação contida em cada componentes e seus antecessores, assim chegou-se no valor mínimo de componentes para se obter, pelo menos, 80% da variablidade dos dados. A Tabela 4.1 mostra a variância de cada componente, e pode-se ver que são necessários 20 componentes para se alcançar 80% de variabilidade.

Tabela 4.1: Variância das componentes

Componente	Variância	Variância acumulada
1	0.2108	0.2108
2	0.0801	0.2910
3	0.0629	0.3539
4	0.0510	0.4050
5	0.0465	0.4515
6	0.0402	0.4917
7	0.0366	0.5284
8	0.0349	0.5633
9	0.0273	0.5907
10	0.0253	0.6161
11	0.0244	0.6405
12	0.0231	0.6636
13	0.0216	0.6853
14	0.0194	0.7047
15	0.0186	0.7233
16	0.0172	0.7406
17	0.0168	0.7574
18	0.0165	0.7739
19	0.0163	0.7903
20	0.0155	0.8058

Com apenas 33,3% das variáveis consegue-se uma variância de 80,58% dos dados tendo uma redução da dimensionalidade de 66,7% mantendo mais de 80% da variância. Na Figura 4.1 tem-se o gráfico das variâncias e seus valores.

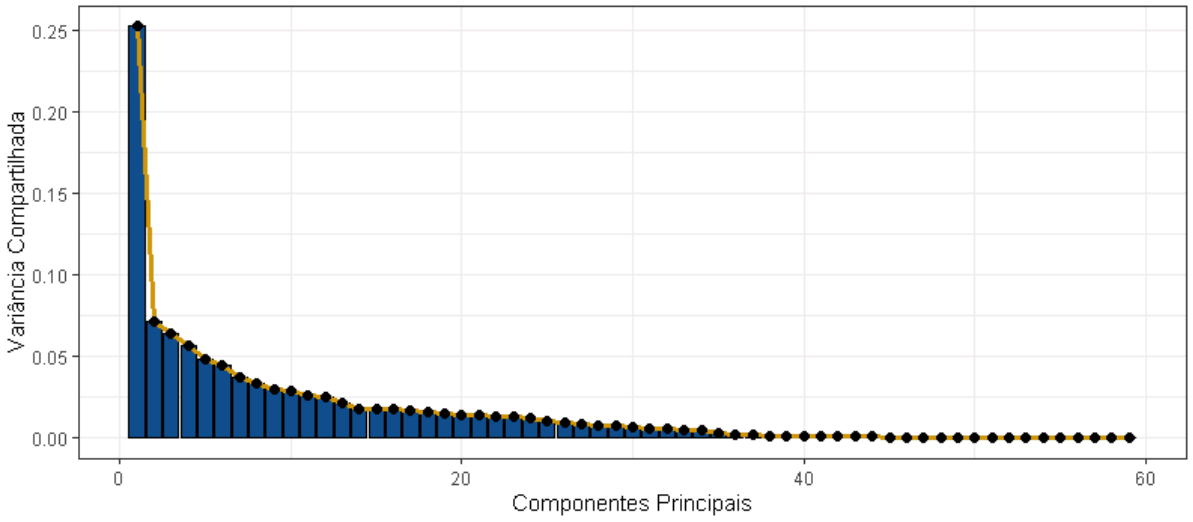

Figura 4.1: Variancias acumulada das componentes

Com as componentes principais definidas e escolhidas a próxima etapa é a construção do modelo de regressão logística, onde as variáveis explicativas são as 20 componentes definidas ateriormente. O modelo apresentou as seguintes características:

Tabela 4.2: Informações da regressão com componentes principais

Estatística	Valor
Acurácia	0.6669
Sensitividade	0.6811
Especificidade	0.5001
Valores preditos positivos	0.9411
Valores preditos negativos	0.1179
K-S	0.2388
ROC	0.6508
Pseudo R^2	0.0510

Após o modelo criado tem-se a matriz de confusão e a equação da regressão realizada com componentes principais:

		Valor predito	
		0	1
Valor observado	0	140555	8793
	1	65814	8795

Equação resultante da regressão logística utilizando as componentes principais

$$Y = -7.77215 + 0.22713PC1 + 0.23376PC2 + 0.05575PC3 + 0.17316PC4 - 0.38639PC5 + 0.02655PC6 + 0.06180PC7 + 0.22863PC8 - 0.13134PC9 - 0.12414PC10 + 0.04170PC11 + 0.26587PC12 - 0.07110PC13 - 0.05193PC14 + 0.25643PC15 + 0.09713PC16 - 0.16538PC17 + 0.08516PC18 - 0.017888PC19 - 0.11523PC20 \quad (4.1)$$

As variáveis abaixo são as variáveis do modelo já usado pelo BRB, sendo modelado através de regressão logística, com suas variáveis categorizadas:

Tabela 4.3: Informações do modelo de regressão usado pelo BRB

Variavel	Pesos
Valor do Score da Serasa	VLSCORESERASA
Qtde Operações Vencidas(SFN)	OCSQTDVENC
Vlr Total Modalidade 19(SFN)	OCSVALM19
Saldo Médio Poupança	SMPSLDAPA
Cód. Segmaneto Cliente	SGMCODSEG
Grau de Instrução	FISESTINSTR
Saldo Médio CDB	SMASSLDFCB

Abaixo a matriz de confusão e informações do modelo de regressão logística utilizado pelo BRB:

		Valor predito	
		0	1
Valor observado	0	21.664	127.684
	1	4.129	70.480

Estatísticas do modelo do BRB, regressão logística. Devido ao sigilo das regras negociais do BRB não pode-se incluir a equação da regressão logística que define o modelo. O modelo classifica o indivíduo em uma escala de 1 a 9 segundo sua probabibilidade de *default*, sendo 1 a melhor categoria e 9 a pior. Será considerado *default* o indivíduo com classificação menor, ou igual a 3.

Tabela 4.4: Informações do modelo de regressão usado pelo BRB

Estatística	Valor
Acurácia	0.4114
Sensitividade	0.8399
Especificidade	0.3557
Valores preditos positivos	0.1451
Valores preditos negativos	0.9447
K-S	0.1348
ROC	0.5663
PseudoR^2	0.1491

No ambiente de risco de crédito o objetivo maior é evitar os maus pagadores,

ou seja, evitar classificar clientes como bons, mas na verdade são maus pagadores. No modelo de regressão do BRB o valor do evento de interesse é trocado, para que quanto maior a probabililidade de *default* maior seja a probabilidade do cliente ser um bom pagador. Para o erro do tipo I o modelo do BRB tem 127.684 clientes bons pagadores, classificados como ruins. Para o modelo de componentes principais temos 8793 clientes bons que foram classificados como ruins. O segundo erro é classificar o cliente como bom, mas ele é ruim. Para o modelo do BRB tem-se 4.129 clientes que são ruins, porém foram classificados como bons. Para o modelo de componentes principais tem-se 65.814 clientes que são ruins, mas foram classificados com bons. Para o crédito o segundo erro é o menos impactante pois há 9 classificações em que o cliente pode se enquadrar, com isso ao invés de conceder ou não crédito apenas restringe-se alguns produtos e oferta-se outros calibrando o modelo através das 9 classificações possíveis. O modelo de componentes principais acertou 140.555 clientes adimplentes, contra 21.664 do modelo de regressão clássica, totalizando 118.891 clientes bons que não seriam negados. O modelo de componentes principais acertou 8.795 clientes inadimplentes contra 79.489 clientes do modelo de regressão clássica, totalizando 61.685 clientes ruins que seriam aceitos. Com o modelo de ACP teria-se aceito 118.891 clientes bons e aceito 61.685 clientes ruins totalizando um saldo líquido de 57.206 clientes bons. Esse montante equivale a R$ 1.410.740.566,00 em limite de crédito que seria concedido para esses clientes bons.

Capítulo 5

CONCLUSÃO

Para o modelo de componentes principais nota-se uma boa redução da dimensionalidade, onde tinha-se 60 variáveis e reduziu-se para apenas 20. Mesmo com essa redução de 66,6% ainda tem-se mais de 3 componentes o que dificulta e até inviabiliza a explicabilidade do modelo. Notas-se um ganho de mais de 10% no indicador K-S entre um modelo e outro. Para o modelo de regressão logística tem-se um K-S de 13,48% e para o modelo com componentes principais 23,88% um bom valor para modelos de *Application* para concessão de crédito. O K-S é o indicador de maior relevância no ambiente de risco de crédito e para esse estudo houve um alto ganho entre o modelo tradicional e o novo modelo de componentes principais.

Para o desenvolvimento do modelo de forma rápida e utilizando uma amostra pequena, em relação a base completa, o ganho foi bastante significante sendo o aumento do K-S de 10,4% sendo um valor comparado com modelo de *Machine Learning* com um alto grau de complexidade.

Referências Bibliográficas

Agresti, A. (2015). *A Introdution to Categorical Data Analysis*, (3th ed.). Wiley.

ANBC (2021). Empresas que compõem a ANBC. Disponível em: URL `https://anbc.org.br/`. Acesso em: 07 de mar. de 2021.

Casella, G. & Berger, R. L. (2018). *Inferência Estatística*, (2nd ed.). Cenage Learning.

Exame (2021). Sobre o surgimento da onda digital. Disponível em: URL `https://exame.com/revista-exame/na-onda-digital/`. Acesso em: 07 de mar. de 2021.

Hardle, W. K. & Simar, L. (2015). *Applied Multivariate Statistical Analysis*, (4th ed.). Springer.

Hosmer, D. W. & Lemeshow, S. (2000). *Applied Logistic Regression*, (2nd ed.). Wiley.

Jappelli, T. & Pagano, M. (2000). Information sharing in credit markets: A survey. *Working paper*, 36:8–10.

Johnson, R. A. & Wichern, D. W. (2002). *Applied Multivariate Statistical Analysis*, (5th ed.). Pearson.

Rencher, A. C. (2002). *Methods of Multivariate Analysis*, (2nd ed.). John Wiley and Sons, INC.Publication.

SERASA (2021). Sobre o cenário econômico. Portal SERASA Experian. Disponível em: URL `https://www.serasaexperian.com.br/conteudos/credito/cenario-covid-19-como-ter-seguranca-na-decisao-de-credito/`. Acesso em: 07 de mar. de 2021.

Sicsú, A. L. (2009). *Credit Scoring, Desenvolvimento - Implantação - Acompanhamento*. Blucher.

Capítulo 6

APÊNDICE

6.1 Variáveis utilizadas no modelo de PCA

Apêndice1: Variáveis completas utilizadas no modelo de componentes principais:
INFQTDOPE – Qtd Total de Operações no SFN (SFN/SCR-BACEN);
MFCSLDMMDCC – Saldo Médio Mensal em Conta-Corrente do Cliente;
MFCVALMOV – Média do Valor Mensal de Movimentação Financeira (R$);
QTDMESES – Tempo de Abertura de Conta, em Meses;
VLSCORESERASA – Pontuação do Score recuperado no SERASA (PF e PJ);
CNSQTDSPF – Qtd Pendências Financeiras (SERASA);
MAXVLPAGAMENTOS6M – Maior valor na categoria "Pagamento", últimos 6 meses;
MEANVLDEBCONTA6M – Média das transações "Deb Conta CNP", últimos 6 meses;
MAXVLTRANSFERENCIA6M – Maior valor na categoria "Transferencia", últimos 6 meses;
MEANVLDEPOSITO6M – Média das transações "Deposito", últimos 6 meses;
MAXVLCREDITOPIX6M – Maior valor das transações "CREDITO PIX", últimos 6 meses;
MEANVLDEBITOPIX6M – Média valor das transações "DEBITO PIX", últimos 6 meses;
MEANQTDCREDITOPIX6M – Média da quantidade transações, "CREDITO PIX", últimos 6 meses;
MEANQTDDEBITOPIX6M – Média da quantidade transações "DEBITO PIX", últimos 6 meses;
MAXVLCOBRANCABRB6M – Maior valor transações "PAGAMENTO CO-

BRANCA BRB", últimos 6 meses;
MAXQTDSAQUE6M – Maior quantidade transações "Saque", últimos 6 meses;
MEANVLCOMPRAS6M – Média transações "Compras", últimos 6 meses;
MAXVLGASTOC6M – Maior valor transações "Deb Conta CNP", "Pgtos"e "Comp", últimos 6 meses;
QTSALDO1006M – Qtd meses com saldo, excluindo limite, últimos 6 meses;
VLCARTAO30DAP – Saldo a vencer até 30 dias do cartão, SCR; CNSQTDSRT – Qtd Restrições Financeiras (SERASA);
CNSQTDSPT – Qtd Protestos (SERASA);
OCSQTDM19 – Qtd Operações na Modalidade 19 (SFN/SCR-BACEN); CNSQTDSAJ – Qtd Ações Judiciais SERASA;
INFQTDIFS – Qtd Total de Instituições Financeiras (SFN/SCR-BACEN);
OCSVALM19 – Valor Total de Operações na Modalidade 19 (SFN/SCR-BACEN);
MEANVLPAGAMENTOS6M – Média transações "Pagamentos", últimos 6 meses;
MAXVLSAQUE6M – Maior valor transações "Saque", últimos 6 meses;
MEANVLTRANSFERENCIA6M – Média valor transações "Transferencia", últimos 6 meses;
MAXVLTED6M – Maior valor transações "TED", últimos 6 meses;
MEANVLCREDITOPIX6M – Média valor transações "CREDITO PIX", últimos 6 meses;
QTDCREDITOPIX6M – Qtd transações "CREDITO PIX", últimos 6 meses;
QTDDEBITOPIX6M – Qtd transações "DEBITO PIX", últimos 6 meses;
MAXVLTITULOBRB6M – Maior valor transações "Pgto de tit. do BRB", últimos 6 meses;
MEANVLCOBRANCABRB6M – Média valor transações "Pgto cobr. BRB", últimos 6 meses;
MEANQTDSAQUE6M – Média qtde transações "Saque", últimos 6 meses;
MAXVLGASTO6M – Maior valor transações "Deb Conta CNP", últimos 6 meses;
MEANVLGASTOC6M – Média valor transações "Deb Conta CNP"e "Pgto", últimos 6 meses;
MAXFAIXAATRASO1006M – Maior faixa atraso acima de R$ 100,00, últimos 6 meses;
PCTCREDNEGATIVOOUTROS360D – Valor a vencer nos próximos 360 dias divido pelo saldo a vencer total;
SMASLDFCB – Saldo Médio Mensal em CDB;
QTDPROTATV – Qtd de protesto ativos SERASA;
SMPSLDAPA – Saldo Médio Mensal em Poupança;

CLTDATNAS – Idade do Cliente (em anos);
OCSVALM2 – Valor Total de Operações na Modalidade 2 (SFN/SCR-BACEN);
VLPROTATV – valor dos protestos ativos SERASA;
MAXVLDEBCONTA6M – Maior valor transações "Deb Conta CNP", últimos 6 meses;
MEANVLSAQUE6M – Média valor transações "Saque", últimos 6 meses;
MAXVLDEPOSITO6M – Maior valor transações "Deposito", últimos 6 meses;
MEANVLTED6M – Média valor transações "TED", últimos 6 meses;
MAXVLDEBITOPIX6M – Maior total transações "DEBITO PIX", últimos 6 meses;
MAXQTDCREDITOPIX6M – Maior qtd transações "CREDITO PIX", últimos 6 meses;
MAXQTDDEBITOPIX6M – Maior qtd transações "DEBITO PIX", últimos 6 meses;
MEANVLTITULOBRB6M – Média valor transações "Pgto de titulo do BRB", últimos 6 meses;
QTDSAQUE6M – Qtd transações "Saque", últimos 6 meses;
MAXVLCOMPRAS6M – Maior valor transações "Compras", últimos 6 meses;
MEANVLGASTO6M – Média valor transações "Deb Conta CNP"e "Pgto", últimos 6 meses;
QTSALDO6M – Qtd meses com saldo, nos últimos 6 meses;
QTDMESESSEPAGADORAP – Qtd meses saldo a vencer acima de R$ 100,00.