UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DEPARTAMENTO DE ESTATÍSTICA

Aplicação de Modelos de Crédito Utilizando Pacotes do R

Autor: Tiago Luigi Guadagnin Radin
Orientadora: Profª Drª *Lisiane Priscila Roldão Selau*

Porto Alegre, 2023

‘

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DEPARTAMENTO DE ESTATÍSTICA

Aplicação de Modelos de Crédito Utilizando Pacotes do R

Autor: Tiago Luigi Guadagnin Radin

Trabalho de Conclusão de Curso
apresentado para obtenção do
grau de Bacharel em Estatística.

Banca Examinadora:

Prof. Dra. Lisiane Priscila Roldão Selau, UFRGS

Bel. Rafaela Vidal Galetto, Banco Agibank

Porto Alegre, 2023

‘

“Mais cedo ou mais tarde sempre chegará o dia em que teremos a certeza de que não foi em vão termos feito, sempre que possível, um pouco além daquilo que era nosso estrito dever”.

Cyro Dutra Ferreira

Em busca da realização dos sonhos.

‘

Agradecimentos

Agradeço à UFRGS e ao seu Instituto de Matemática e Estatística pelo ensino e pelo crescimento pessoal e profissional.

Agradeço a professora Lisiane Selau, pelas orientações, por acreditar e por confirmar em mim durante esse processo.

Agradeço aos colegas de trabalho pela oportunidade de ingressar no mercado de trabalho e por toda a ajuda com meu crescimento.

Agradeço a todos os amigos de longa data e aos colegas do curso pela ajuda e parceria ao longo de todo o período.

Agradeço aos meus avós Luigi e Rosina, em memória, e Edmundo e Fidelia por terem sido pessoas a frente do seu tempo e terem perpetuado isso em seus filhos, netos e bisnetos.

Agradeço aos meus pais, Agostinho e Firléia, por sempre terem me auxiliado ao longo de todos esses anos. Agradeço também ao meu irmão, Augusto, pela parceria de sempre. Vocês são inspiração, base e alavanca para mim!

Agradeço a minha eterna namorada, Paola, por aceitar dividir comigo os seus dias, a sua companhia e sua história comigo. Por alegrar a minha caminhada e por ser um exemplo e uma inspiração para eu ser cada dia uma pessoa melhor. Amo você para sempre!

Agradeço ao maior presente e a maior surpresa que essa vida me deu, ao meu melhor amigo, ao meu filho, Mathias. Obrigado por cada sorriso e abraço, combustíveis esses que movem o mundo. A pessoa que sou hoje se deve muito a tua chegada, estarei sempre contigo. Te amo muito!

‘

Resumo

O mercado de crédito no Brasil vem crescendo constantemente ao longo dos últimos anos, dessa maneira, é essencial saber quais clientes irão pagar seus créditos, para mitigar gastos com inadimplentes e assim poder oferecer melhores ofertas de crédito aos bons pagadores. Atualmente são utilizados modelos de pontuação de crédito para auxiliar nessa decisão e com o avanço da tecnologia os softwares de programação ganharam espaço importante para a construção dos modelos. O *software R* é hoje um dos principais programas utilizados para realizar análises de *credit scoring* e nele constantemente são adicionados códigos para auxiliar nessa tarefa. Assim, o presente trabalho tem como objetivo apresentar alguns dos pacotes disponíveis no R e suas funções e quais seriam indicados para percorrer todas as etapas de construção de um modelo de crédito. Inicialmente é realizada uma revisão da literatura, mostrando a importância do *software R* para os modelos de pontuação de risco de crédito. Na sequência é apresentado um exemplo prático de modelo, passando por todas as etapas de construção do mesmo: (i) Delimitação da População; (ii) Seleção da Amostra; (iii) Análise Preliminar; (iv) Construção do Modelo; (v) Escolha do Modelo; (vi) Passos para Implantação. Por fim, são expostas todas as funções e pacotes utilizados para a realização do exemplo prático. Conclui-se que existem diversas peculiaridades para implementar um modelo de pontuação de risco de crédito e que, além de programar, é necessário conhecer a empresa e o mercado que se está querendo modelar, a fim de encontrar os melhores resultados.

Palavras-chaves: *Software R*; Risco de Crédito; Modelo de Pontuação de Risco de Crédito.

‘

Abstract

The credit market in Brazil has been growing steadily over the last few years, so it is essential to know which customers will pay their credits to mitigate expenses with defaulters and thus be able to offer better credit offers to good payers. Currently, credit assessment models are used to assist in the decision and with the advancement of technology, programming software has gained important space for the construction of models. The R software is today one of the main programs used to perform credit score analysis and codes are constantly added to auxiliary tasks in this task. Thus, the present work aims to present some of the packages available in R and their functions and which ones would be indicated to go through all the stages of building a credit model. Initially, a literature review is carried out, showing the importance of the R software for credit risk assessment models. Next, a practical example of a model is presented, going through all the stages of its construction: (i) Population delimitation; (ii) Sample Selection; (iii) Preliminary Analysis; (iv) Model Construction; (v) Choice of Model; (vi) Steps for Implementation. Finally, all the functions and packages used to carry out the practical example are exposed. It is concluded that there are several peculiarities to implement a credit risk scoring model and that, in addition to programming, it is necessary to know the company and the market that one is trying to model, in order to find the best results.

Keywords: *Software R*; Credit Risk; Credit Risk Scoring Model.

‘

Sumário

‘

1. Introdução

O crédito está presente no cotidiano de todos. Ao fazer um planejamento qualquer, pensando na relação entre o que se quer obter e os recursos finitos existentes, está-se aplicando o princípio de crédito na vida real (SCHRICKEL, 1997, p. 11). Para as instituições financeiras que emprestam crédito, um dos principais dilemas está envolto na incerteza de saber se haverá o retorno do crédito emprestado, sendo isso definido como o risco de crédito, segundo Duarte Júnior (1996, p. 3).

Técnicas da estatística podem contribuir para estimar e analisar modelos de crédito, reduzindo a incerteza intrínseca dessa área. Ferramentas e softwares são desenvolvidos para auxiliar e continuam obtendo atualizações constantes, para facilitar esse processo e oferecer resultados mais significativos (SZEPANNEK, 2020). Os modelos de crédito são ferramentas amplamente utilizadas por bancos e financeiras para medir risco em conceder crédito, analisando dados de quem solicitou, com a finalidade de aumentar a receita e diminuir os prejuízos das empresas, focando a liberação de crédito aos bons clientes. Neste contexto, o tema central desta pesquisa é a modelagem de crédito.

Dada a demanda de melhorar as modelagens de crédito, diversos pacotes e funções foram criados para ajudar na realização dessas medições. Porém, com o grande número de possibilidades, acabam surgindo dúvidas de qual o melhor pacote utilizar, como utilizar determinadas funções e como fazer a análise dos resultados. Em se tratando do *software* R, percebe-se que mesmo tendo surgido anos depois que outros programas, como o SAS, que contém diversas soluções para a modelagem de crédito, o *software* R apresenta uma grande quantidade de novos pacotes disponíveis aos usuários, que também se multiplicaram de maneira significativa, tornando este programa um dos mais utilizados para este tipo de análise. Pacotes do R servem para organizar as funções que darão o retorno esperado no R (MAYER E ZEVIANI, 2016), que visam facilitar o trabalho do cientista. Funções, por sua vez, são códigos pré-programados que executam uma sequência de tarefas.

Por esta razão, nesta pesquisa será feito o estudo de alguns pacotes disponíveis no *software* R para apresentar diferentes maneiras de construção de modelos de crédito, exemplificando as principais etapas do processo, desde a seleção da amostra até a escolha do modelo, por exemplo.

‘

Buscando facilitar o trabalho dos cientistas de dados das instituições financeiras que estão incumbidos de construir modelos de crédito utilizando o *software* R, o objetivo principal deste trabalho é apresentar alguns dos pacotes disponíveis e suas funções e quais seriam indicados para percorrer todas as etapas de construção de um modelo de crédito. Dessa forma, serão apresentados elementos para que os usuários do *software* R no ambiente acadêmico e em instituições do mercado financeiro possam realizar a análise dos riscos na concessão de crédito com mais facilidade.

2. Referencial Teórico

2.1 Risco de Crédito

O crédito para um banco pode ser exemplificado como quando se coloca uma quantia de dinheiro disponível para um indivíduo fazer uso, mediante um compromisso de que o mesmo fará o pagamento em data futura. Quando um contrato é firmado, cria-se o planejamento de que o valor será pago com acréscimo de juros no tempo estipulado. Assim, segundo Brito e Assaf Neto (2008), o risco de crédito é a possibilidade de o tomador gerar perdas, através das obrigações que não cumpriu.

O risco para o credor passa por uma escala de 0 a 100% gradativamente, baseando-se em variáveis quantitativas e qualitativas, que alteram o *rating* do tomador (BRITO et al., 2009). Segundo Neves et al. (2007), o *rating* tem em sua essência um caráter informativo. Conforme a chance de o cliente não pagar sua dívida aumenta, aumentando os dias de atraso, por exemplo, a qualidade de bom pagador vai sendo impactada negativamente, influenciando em indicadores como prazo, taxa de juros e quantia de empréstimo.

2.2 Modelos de Pontuação de Risco de Crédito

Os modelos de pontuação de risco de crédito servem principalmente para analisar quanto risco um indivíduo ou uma carteira oferecem à instituição bancária. Os modelos atribuem uma medida normalmente chamada de *rating* ou escore, que influenciará nos indicadores do empréstimo citados na seção anterior (BRITO E ASSAF NETO, 2008). Tais modelos de crédito são comumente chamados de *Credit Scoring*.

Segundo Bacconi e Gouvêa (2005), 90% das empresas americanas que oferecem algum tipo de crédito ao consumidor já utilizavam modelos de *credit scoring*, porém no Brasil esse

‘

uso em grande escala começou apenas em meados dos anos 90. Quando se trata de um banco, com milhares de clientes novos todos os meses, os modelos permitem tomar decisões rápidas para se manter competitivo no mercado, classificando as novas solicitações a partir dos dados que o analista informa ao sistema, que retorna positiva ou negativamente para a aprovação do novo contrato (SELAU, 2009).

Segundo Selau (2009), os métodos para modelos de crédito mais empregados são as técnicas econométricas, que compreende conhecimentos estatísticos como a regressão logística, as técnicas de redes neurais, que busca uma solução através de um processo de aprendizagem, as técnicas de modelos de otimização, que buscam os pesos ideais das variáveis para maximizar os lucros, os sistemas especialistas, que se baseiam em regras decisórias na tomada de decisão, e os sistemas híbridos, que utilizam técnicas de estimativa e simulações diretas.

2.3 *Software* R para Modelos de Crédito

O *software* R contém uma linguagem de programação com código aberto e que vem tendo uma grande expansão nos mercados acadêmico e financeiro, por ser uma linguagem mais flexível que apresenta uma extensa coleção de pacotes que os pesquisadores desenvolvem e disponibilizam no site da linguagem R. Owen (2010) classifica o programa como um conjunto de recursos para manipulação de dados, permitindo fazer simulações, cálculos e até exibições gráficas de maneira eficaz.

Segundo Sharma (2009), até meados de 2009 não existiam guias ou documentações sobre modelos de crédito no R, dessa maneira pode-se considerar como recente o desenvolvimento de materiais como esse, visando aproximar a ferramenta do público em geral. Assim como Szepannek (2020), os pacotes serão apresentados no presente trabalho entre as etapas a serem seguidas, apresentando os pontos fortes e fracos de cada função descrita.

3. Metodologia de Pesquisa

A ideia principal deste trabalho é aprofundar conhecimentos acerca dos pacotes do *software* R utilizados para construir modelos de crédito. Assim, segundo Gerhardt e Silveira (2009), trata-se de uma pesquisa qualitativa. Ainda segundo os mesmos autores, esta será uma pesquisa básica exploratória, uma vez que se quer gerar novos conhecimentos aos usuários, proporcionando maior familiaridade na resolução do problema em estudo. Foi necessário

‘

realizar uma pesquisa bibliográfica, baseada em trabalhos científicos disponíveis referente ao uso de pacotes do *software* R.

O estudo foi desenvolvido através das etapas descritas a seguir:

1. Buscar bibliografia que apresenta diferentes opções de pacotes para realizar a modelagem de crédito. Assim sendo, os dados desta pesquisa são os pacotes e suas funções.

2. Organizar os dados obtidos, detalhando em qual parte do processo de modelagem eles se encaixam e quais suas funcionalidades.

3. Seguir os passos e a metodologia proposta por SELAU (2009, p. 66) para realizar um Modelo de Previsão de Risco de Crédito: dividir e exemplificar os pacotes e suas funções, que desempenham cada etapa do processo, composto por:

a) Delimitação da População: compreende escolher qual a população a ser estudada, a existência de dados da população e a definição do desempenho em satisfatório e insatisfatório;
b) Seleção da Amostra: compõe fazer a seleção de um grupo representativo dentro da população escolhida, além da identificação de variáveis disponíveis; realizar uma validação dos dados e separar a amostra para análise e teste;
c) Análise Preliminar: tem como objetivo escolher dentre as variáveis disponíveis quais devem entrar na modelagem e a criação de variáveis *dummies*: aquelas que tomam o valor de “zero” ou de “um”.
d) Construção do Modelo: abrange escolher a técnica que será usada, da seleção das variáveis independentes e da verificação se as suposições que norteiam a técnica são atendidas;
e) Escolha do Modelo: engloba avaliar o percentual de classificações corretas, além do teste KS para duas amostras e da área sob a curva ROC;
f) Passos para Implantação: envolve programar a implementação junto a empresa e quais são os pontos a serem observados para o melhor desempenho do modelo. Esta parte não será apresentada no capítulo seguinte por não conter funções do *Software* R nesse trabalho.

‘

4. Desenvolver material de apoio com o resumo das funções utilizadas para construir o Modelo de Previsão de Risco de Crédito, trazendo os principais resultados observados em cada etapa do processo.

5. Discutir os resultados encontrados, apontando observações que devam ser feitas ao desenvolver um Modelo de Previsão de Risco de Crédito.

4. Aplicação

Ao longo deste capítulo serão apresentadas as funções utilizadas para realizar um exemplo de construção de um modelo de pontuação de crédito, e seus respectivos pacotes no *software* R (R CORE TEAM, 2011). Diferentes análises podem ser feitas conforme o banco de dados em estudo, buscando alcançar os objetivos desejados pelo cientista de dados. Algumas funções que serão apresentadas fazem parte do pacote básico do *software* R.

4.1 Delimitação da População

Como apresentado na metodologia, nesta etapa da construção do modelo é preciso decidir qual segmento da população será estudada para obter a sua pontuação de crédito. Para isso, segundo Gouvêa e Gonçalves (2006), é necessária uma base de dados com qualidade e disponibilidade e que os clientes mantenham um comportamento similar ao longo do tempo.

Para o presente trabalho, para exemplificar os pacotes e funções do *software* R, será utilizado um banco de dados de uma rede de farmácias com unidades em todo o Rio Grande do Sul, para os quais é oferecido um cartão de crédito próprio a fim de facilitar o pagamento das compras. Os cadastros foram realizados de dezembro de 2005 até julho de 2006 e as informações pessoais foram transformadas para preservar o sigilo dos dados.

Escolhida a população a ser estudada e tendo a base de dados, ainda é preciso definir o desempenho desejado pelo concessor quanto a satisfatório e insatisfatório. Para a empresa em questão o bom pagador é aquele que quita suas dívidas com no máximo 30 dias de atraso e os maus pagadores são aqueles com atraso superior a 60 dias. Devem ser removidos do banco de dados clientes com atrasos entre 31 e 60 dias e clientes que não utilizaram o cartão por apresentarem resultado indeterminado entre as categorias desejadas.

‘

4.2 Seleção da Amostra

Nesta etapa do processo será utilizada a primeira função do R. Trata-se da função *read_excel*, do pacote *readxl*, que será utilizada para carregar o banco de dados no *software* R, com a seguinte sintaxe:

dados = *read_excel*("<nome_do_arquivo>.xlsx") (1)

É necessário realizar uma validação dos dados disponíveis, a fim de conhecer as variáveis, assim como procurar transformações que possam ser feitas, além de remover de maneira prévia variáveis que não contribuirão para a construção do modelo. As variáveis disponíveis no banco de dados escolhido estão apresentadas na Figura 1:

Variável	Definição
cod_cli	Código do cliente
sexo	F - Feminino ou M - Masculino
uf_natu	UF da cidade de naturalidade
est_civ	Estado civil
dtnasc	Data de nascimento (dd-mmm-aaaa)
tp_resid	Tipo de residência
gra_inst	Grau de instrução
tp_ocup	Tipo de ocupação
cep_com	CEP comercial (xxxx-xxx)
dt_admi	Data de admissão no emprego (dd-mmm-aaaa)
flg_pens	Pagamento de pensão alimentícia (S - sim ou N - não)
flg_filh	Tem filhos (S - sim ou N - não)
dtcad	Data de cadastro
cred_3s	Possui cartões de outros comércios (S - sim ou N - não)
tp_salar	Tipo de salário (D - declarado ou C - comprovado)
tipo	1 - Bom pagador e 0 - Mau pagador

Figura 1- variáveis do banco de dados utilizado e suas descrições

‘

A variável tipo é que identifica os bons e maus pagadores, será realizada uma contagem de quantos indivíduos o banco possui em cada categoria, utilizando a função *count*, do pacote *dplyr*, com a seguinte sintaxe:

```
dplyr::count(dados, “tipo”)
```
(2)

O resultado da sintaxe acima será uma tabela mostrando que o banco de dados possui 4.589 linhas de maus pagadores, representados pela variável tipo igual a 0 e 6.958 linhas de bons pagadores, representados por 1, totalizando assim 11.547 indivíduos que serão analisados nesse exemplo, como apresentado na Tabela 1:

Tabela 1 - Frequência de bons e maus pagadores

Tipo	Frequência
0 - Mau	4.589
1 - Bom	6.958

Dado que a base de dados apresenta a data de nascimento e a data de cadastro, pode-se calcular a idade de cada indivíduo quando teve acesso ao cartão de crédito, utilizando a função *difftime*, com a seguinte sintaxe:

```
idade_cad = as.numeric(difftime (dados$dtcad, dados$dtnasc,
        units = "weeks"))/52.25
```
(3)

É importante ressaltar que com outros bancos de dados, podem ser feitas outras criações de variáveis para auxiliar o cientista de dados a obter na modelagem o melhor resultado possível. Outra informação importante para ser analisada é a quantidade de erros ou *missings* presentes nas variáveis. Essa análise pode ser feita com a função *sapply*:

```
sapply(dados, function(x) sum(is.na(x)))
```
(4)

Dessa maneira pode-se ver que a variável dt_admi possui 8.421 dados faltantes e a variável cep_com, 5.272. Assim, essas variáveis serão descartadas da análise, sendo necessário o cuidado futuro para que sejam de preenchimento obrigatório para todos os clientes. Será removida também a coluna cod_cli por ser apenas um código que identifica cada cliente, além das variáveis dtcad e dtnasc por já terem sido utilizadas na criação da variável idade_cad, utilizando a seguinte função:

‘

```
dados = dados[,-c(which(names(dados) %in% c("cod_cli", "cep_com",
        "dt_admi", "dtcad", "dtnasc")))]
```
(5)

Além disso, devem ser excluídas as linhas que apresentam dados faltantes, com função *na.omit*, feita através da sintaxe abaixo. Dessa maneira foi realizada a primeira validação, onde o conjunto de dados ficou contendo 11.390 linhas, pouco menor que a quantidade original.

```
dados = na.omit(dados)
```
(6)

Dado que a única variável numérica é a idade_cad, todas as demais serão convertidas em fatores, para poder incorporar no modelo de crédito. Para isso, foi utilizado o operador *pipe* (%>%) do pacote *magrittr* e a função *mutate_at*, do pacote *dplyr*, com a seguinte sintaxe:

```
fatores = names(dados)[!names(dados) %in% c("idade_cad")]
dados = dados %>% mutate_at(fatores, list(~factor(.)))
```
(7)

Com esses passos concluídos, o banco está pronto para seguir para o próximo passo da construção do modelo de pontuação de crédito.

4.3 Análise Preliminar

Nessa etapa pode-se começar visualizando as variáveis. Para o caso das numéricas, como idade_cat, pode-se usar a função *plot_boxplot*, do pacote *DataExplorer* para visualizá-la por tipo de pagador, utilizando a seguinte sintaxe:

```
plot_boxplot(dados,by="tipo")
```
(8)

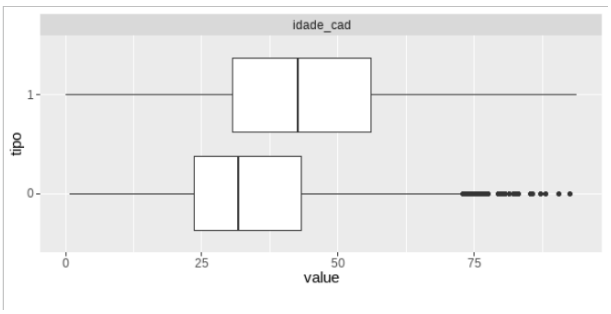

Figura 2 - Boxplot da idade no cadastro para bons (1) e maus pagadores (0).

Na da Figura 2, é possível perceber que a idade se distribui para valores maiores entre os clientes bons pagadores. Já para as variáveis que são fatores, como sexo por exemplo, pode-

‘

se utilizar a função *ggplot*, do pacote *ggplot2*, que apresenta diferentes combinações, como pode ser visto na Figura 3:

```
ggplot(dados, aes(y = sexo, fill = tipo)) + geom_bar(position = "fill") +
labs(x = "Proporção", y = "Sexo", fill = "Classificação")
```
(9)

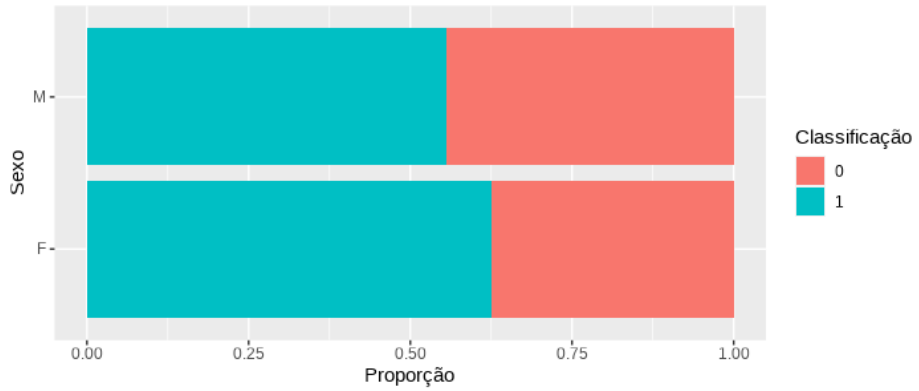

Figura 3 - Histograma com a contagem dos bons e maus pagadores por sexo dos indivíduos.

Para as variáveis categóricas com vários níveis é possível utilizar o processo de categorização, que consiste em agrupar os valores assumidos pelas variáveis em classes de tal forma que seja informativo para a discriminação entre bons e maus pagadores. Será utilizado um processo automático para agrupá-los quando a variável tiver mais que 5 níveis. A função *woebin* do pacote *scorecard* foi a escolhida nessa etapa:

```
bins = scorecard::woebin(dados, y = "tipo", x = c("uf_natu"),
        check_cate_num = F)
```
(10)

A função *woebin* irá gerar vários retornos para serem utilizados, e nesse caso será consultado o retorno *total_iv*, onde através da função *unique* é possível saber qual o Valor da Informação (IV), ou *Information Value* da variável. Segundo Kauffmann (2017), o IV é uma medida importante para avaliar o poder preditivo de determinada variável e pode ser interpretada da seguinte maneira: valores menores que 0,02 para o IV não possuem poder preditivo; valores entre 0,02 e 0,1 possuem baixo poder preditivo; valores entre 0,1 e 0,3 possuem médio poder preditivo; valores entre 0,3 e 0,5 possuem forte poder preditivo; e valores acima de 0,5 possivelmente são valores suspeitos de uma variável que depende diretamente de outra e deve ser descartada do modelo. Analisando as variáveis descritas na sintaxe (10), é possível perceber que a variável uf_natu obteve um IV igual a 0, assim ela será a próxima variável a ser descartada.

‘

Foi mostrado com a sintaxe (8) que médias das idades dos indivíduos na data de cadastro estão deslocadas considerando os bons e maus pagadores. Dessa maneira, a variável será categorizada para usá-la na modelagem, pois Thomas (2000) ressalta a importância de formar grupos homogêneos dentro de cada categoria e heterogêneo entre elas e Gouvêa e Gonçalves (2006) complementam ao enfatizar que agrupando se evita categorias com um número pequeno de observações. Para auxiliar nessa categorização, será utilizado o princípio do risco relativo (RR), pois quanto maior é a diferença dos percentuais de bons e maus pagadores de uma mesma variável, maior será a utilidade dessa variável (SELAU, 2008). Lewis (1992) divide o RR nas seguintes classes: péssimo (RR < 0,50); muito mau (RR entre 0,50 e 0,67); mau (RR entre 0,67 e 0,90); neutro (RR entre 0,90 e 1,10); bom (RR entre 1,10 e 1,50); muito bom (RR entre 1,50 e 2,00) e excelente (RR maior que 2,00). Buscando atender os conceitos apresentados anteriormente, utilizando o conjunto de comandos abaixo para analisar a variável idade_cad:

```
df = dados %>% select(idade_cad, y=tipo)
df$idade_cad <- floor(df$idade_cad)
dsc0 = df %>% pivot_longer(cols = c(1), names_to = "Var",
values_to = "Valor")
dsc <- data.frame()
dsc = dsc0 %>% group_by(Var,Valor)  %>% summarise(Total = n(),
        Bons = sum(y==1), Maus = Total - Bons)
dscTot = dsc0 %>% group_by(Var) %>% summarise(TotalPop = n(),
        BonsTot = sum(y==1), MausTot = TotalPop - BonsTot)
dsc = dsc %>% full_join(dscTot, by = "Var")
dsc = dsc %>% mutate(RR = (Bons/BonsTot)/(Maus/MausTot))                    (11)
dsc$RR[dsc$RR == Inf] = 4
vars = unique(dsc$Var)
plots = list()
for(i in vars){
k = length(unique(df[[i]]))
k = floor(k/3)
plots[[i]] = dsc %>% filter(Var == i) %>%  ggplot(aes(x=Valor, y=RR)) +
        geom_line() + scale_y_continuous(limits = c(0,4)) + theme_bw() +
        labs(x = i, y="RR", title=i) }
gridExtra::grid.arrange(grobs=plots)
```

Com a sintaxe (11), se obtém o RR para cada valor de idade, conforme a Figura 4:

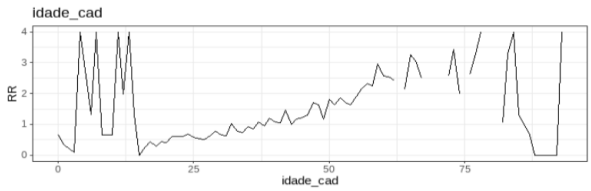

Figura 4 - Gráfico do Risco Relativo para cada valor de idade_cad.

‘

É possível observar uma grande variação no RR até a idade 25 e a partir da idade 65, isso se deve a pequena quantidade de dados que correspondem a esses grupos. Dessa maneira, será feito o agrupamento de 0 a 25 anos, posteriormente criando grupos de 10 anos de diferença, até o último grupo de 65 a 100 anos. Para isso será utilizado o conjunto de funções abaixo:

```
df1 = dados %>% dplyr::select(idade_cad, y=tipo)
kortes = seq(25,65,10)
kortes = c(0,kortes,100)
nomeGrupo = cut(df1$idade_cad, kortes, include.lowest = T, left=T)
df1 = df1 %>% mutate(Grupo = cut(df1$idade_cad, kortes, labels = F,
      include.lowest = T, left=T), nomeGrupo)
df1 = df1 %>%
group_by(Grupo) %>% summarise(Total = n(), Bons = sum(y==1),
      Maus = Total - Bons) %>% mutate(pBons = Bons/sum(Bons),
      pMaus = Maus/sum(Maus), RR = pBons/pMaus)
df1 %>% ggplot(aes(x=Grupo, y=RR)) + geom_line() + theme_bw() +
      geom_point() + scale_x_continuous(breaks = seq(1,6),
      labels = unique(nomeGrupo)) + scale_y_continuous(
      breaks = seq(0,4,0.5), labels = seq(0,4,0.5)) +
      geom_hline(linetype=2, colour="red", yintercept = 1)
```

(12)

Com a sintaxe (12), é obtido o RR para a variável idade agrupada, como na Figura 5:

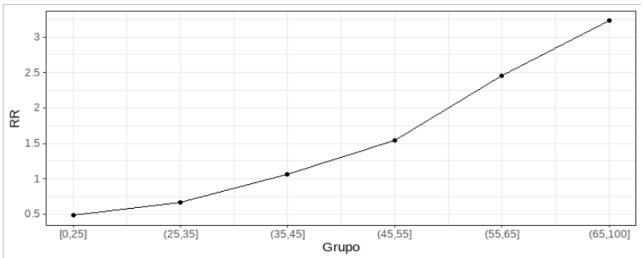

Figura 5 - Gráfico do Risco Relativo para os valores agrupados da idade_cad.

Agora é possível visualizar os critérios de risco relativos necessários para testar se a variável será significativa para o modelo, com grupos de quantidades significativas de dados e cada um heterogêneo com relação aos demais. Assim essa categorização, chamada de idade_bin substituirá a variável idade_cad no banco de dados, que agora está pronto para a etapa de Construção do Modelo.

4.4 Construção do Modelo

Abrange escolher a técnica que será usada, da seleção das variáveis independentes e da verificação se as suposições que norteiam a técnica são atendidas.

‘

Para iniciar a seleção das variáveis que irão compor o modelo será utilizada a função *iv*, do pacote *scorecard*, que retorna o valor de IV de todas as variáveis que compõem o banco após os tratamentos nos passos anteriores, como apresentado na Tabela 2.

```
df=scorecard::iv(dados, y = "tipo", order = T)
```
(13)

Tabela 2 -Valor de Informação das variáveis presentes no banco de dados

variable	info_value
idadeBIN	0.3561876
est_civ	0.1959302
tp_ocup	0.1323415
tp_resid	0.0351044
sexo	0.0186225
gra_inst	0.0102244
flg_filh	0.0094600
tp_salar	0.0030615
cred_3s	0.0009689
flg_pens	0.0007263

Como as variáveis com IV menor que 0,02 não possuem poder preditivo (KAUFFMANN, 2017), elas serão descartadas para o processo de modelagem.

Para avaliar o percentual de classificações corretas do modelo, deve-se separar o banco de dados em um conjunto de treino e um de teste. A função *train_test_split*, do pacote *creditmodel*, permite fazer tal separação com a seguinte sintaxe:

```
traintest = train_test_split(dados, prop = 0.7)
treino = traintest$train
teste = traintest$test
```
(14)

Com a sintaxe (14) foi criada uma base contendo 70% dos dados para treino, que será utilizada para criar o modelo e os demais dados serão usados para testar se o modelo apresentará bons resultados.

Para ajustar o modelo de regressão logística será utilizada a função *glm*, do pacote *stats*, com a qual será obtido o modelo

‘

```
modelo = glm(formula = tipo ~ ., data = treino,
        family = binomial(link = "logit"))
```
(15)

É necessário verificar se existe impacto em razão da multicolinearidade, para que haja significância prática e também estatística (SELAU, 2008). Para isso, será usada a função *check_collinearity*, do pacote *performance*, onde os autores Lüdecke et al. (2021) apresenta que o Fator de Inflação (VIF) não pode ter valores superiores a 10, por apresentar assim uma correlação alta e não tolerável em relação aos preditores do modelo.

```
check_collinearity(modelo)
```
(16)

Como nenhum dos valores foi superior a 10, será aceito o modelo criado na sintaxe (15). Caso se verifique multicolinearidade nos dados, os autores Corrar et al. (2007) apresenta o método de stepwise como uma das possibilidades de ação corretiva do problema, como apresentado na sintaxe (16). A função step, do pacote stats, é uma das possibilidades para resolver o problema (R CORE TEAM, 2011).

```
df = step(dados, scope = 'upper')
```
(17)

4.5 Escolha do Modelo

Nessa etapa será utilizada a base de teste, gerada na sintaxe (14) para comprovar a qualidade do modelo. Com a função *predict*, do pacote *stats*, pode-se testar o modelo, gerado na sintaxe (15), para o banco de dados de treino e de teste criados anteriormente.

```
teste$score<-predict(modelo,type='response',teste)
```
(18)

Para realizar a avaliação do modelo é possível utilizar diferentes ferramentas. Inicialmente será avaliada a curva ROC (receiver operating characteristic), a qual é uma representação gráfica da performance do modelo, cruzando a taxa de verdadeiros positivos no eixo y, ou seja, quando o modelo acertou, pela taxa dos falsos positivos no eixo x, quando o modelo errou. Dessa forma, quanto mais a curva ROC se aproxima do canto superior esquerdo, maior é a qualidade do teste (POLO e MIOT, 2020). Para gerar a curva ROC será utilizada a função abaixo:

```
rocplot <- function(pred, truth, ...) {
```
(19)

‘

```
predob = prediction(pred, truth)
perf = ROCR::performance(predob, "tpr", "fpr")
plot(perf, ...)
area <- auc(truth, pred)
area <- format(round(area, 4), nsmall = 4)
text(x=0.8, y=0.1, labels = paste("AUC =", area))
segments(x0=0, y0=0, x1=1, y1=1, col="gray", lty=2)}
rocplot(teste$score, teste$tipo, col="blue",main="ROC-Teste")
```

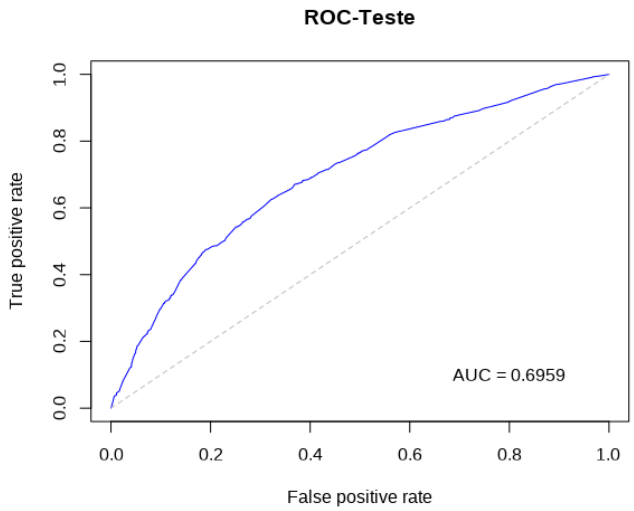

Figura 6 - Curva ROC com os dados de teste.

O valor de AUC indicado na Figura 6, se refere a área debaixo da curva (AUC) e, como o valor ficou maior 0,5, o teste mostra que o modelo está acertando mais do que está errando.

Existe ainda outra maneiras de avaliar a qualidade do modelos, que é o percentual de classificações corretas, mas para isso é preciso definir que os valores preditos com o modelo que forem menores que 0,5 serão considerados maus pagadores, caso contrário, como bons pagadores (SELAU, 2008). A análise é feita cruzando os resultados observados e previstos, calculando a acurácia, que nada mais é do que a proximidade entre o valor obtido experimentalmente no modelo e o valor verdadeiro do banco de dados, e pode ser obtida com a seguinte sintaxe:

```
teste$predito = ifelse(teste$score>=0.5,1,0)
tab_teste = table(teste$tipo,teste$predito)
taxaacerto_teste = (tab_teste[2,2]+tab_teste[1,1])/sum(tab_teste)
print(paste0("Acurácia: ", round(taxaacerto_teste,2)))
```

(20)

‘

Como resultado da sintaxe (20), obtém-se o valor de acurácia 0,66, considerado satisfatório por ser maior que 0,65 (PICININI et al., 2003), apresentado na Tabela 3.

Tabela 3 - Verificações de acertos nas classificações do modelo.

	Previsto – BOM	Previsto – MAU
Observado – BOM	1.608	471
Observado – MAU	685	653

É possível também visualizar a densidade da pontuação obtida no modelo, utilizando novamente o marcador *pipe* e a função *ggplot*.

```
teste %>% ggplot(aes(x=score, fill=factor(tipo))) + geom_density(alpha=0.5)
```
(21)

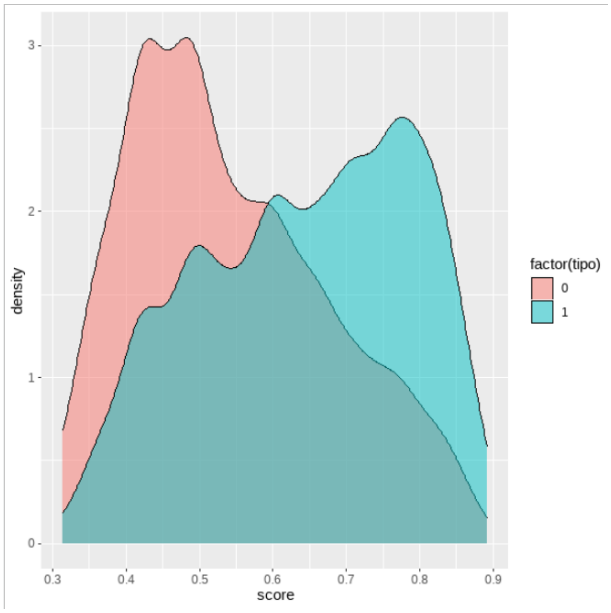

Figura 7 - Distribuição da pontuação de crédito para os maus e bons pagadores

É esperado que os maus pagadores estejam para o lado esquerdo e os bons pagadores no lado direito e, como é possível ver na Figura 7, o modelo consegue fazer essa distinção. Para ajudar ainda mais essa visualização pode-se dividir em decis, onde posteriormente se calcula a taxa de inadimplência, como sendo os maus pagadores pelo total de clientes de cada grupo. Se espera que a linha seja decrescente conforme aumenta o grupo, uma vez que os maus pagadores estão nos grupos mais baixos, além disso, espera-se que não haja linhas crescentes ao longo do gráfico. Isso pode ser feito com a sintaxe abaixo:

```
kortes = quantile(teste$yChapeu, seq(0,1,1/10))
teste = teste %>% mutate(Grupo = cut(yChapeu, breaks=kortes, left=T,
```
(22)

‘

```
        labels=F, include.lowest=T))
dsc = teste %>% group_by(Grupo) %>% summarise(Tot = n(),
        Bons = sum(tipo == 1), Maus = sum(tipo==0)) %>%
        mutate(pMau = Maus/Tot)
dsc %>% ggplot(aes(x=Grupo, y=pMau)) + geom_line() +
        geom_point() + theme_bw() +
        labs(x="Grupo", y="Proporção de Inadimplentes")
```

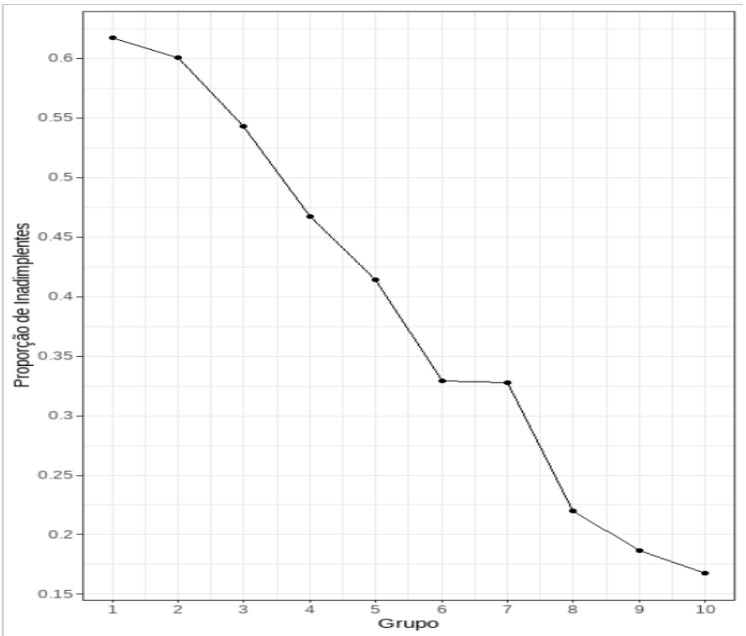

Figura 8 - Inadimplência por grupo de pontuações.

É possível ver na Figura 8 que os critérios apresentados anteriormente foram atendidos. Para finalizar a análise, é possível realizar o teste Kolmogorov-Smirnov (KS) para duas amostras, o qual junto com percentual de classificações corretas é uma das medidas de desempenho mais utilizadas (SELAU, 2008). Para isso será utilizada a sintaxe a seguir:

```
dsc = dsc %>%  mutate(txMau = Maus/sum(Maus), txBom = Bons/sum(Bons),
        acumMau = cumsum(txMau), acumBom = cumsum(txBom),
KS = abs(acumBom - acumMau))
dfPlt = pivot_longer(dsc, cols = c(8,9), names_to = 'Tipo',
        values_to = 'Valores')                                          (23)
dfPlt %>% ggplot(aes(x=Grupo, y=Valores, colour = Tipo)) + geom_line() +
        annotate(geom = 'label',x = 5, y = 0.5,
        label = paste0("KS = ", round(max(dsc$KS),2)))
```

‘

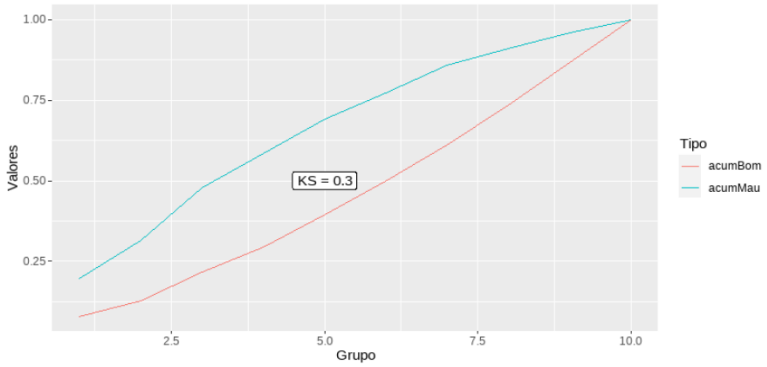

Figura 9 - Teste KS para duas amostras.

Juntando o resultado do teste KS, que atingiu a marca de 30% de diferença, como é possível ver na Figura 9, mostrando também que o modelo é eficiente para prever os bons e maus pagadores (PICININI et al., 2003), aos resultados apresentados anteriormente nos demais testes, pode-se dizer que o modelo atingiu as expectativas para prever a pontuação de crédito ao público utilizado.

5. Material de Apoio

Neste capítulo serão expostas as funções utilizadas no capítulo 4, juntamente com o pacote de origem e uma breve descrição de como foram utilizadas. Serão apresentadas em duas figuras por etapa, uma para as funções disponibilizadas pelos pacotes do R e outra com funções montadas para alcançar a descrição em questão.

5.1 Seleção da Amostra

Na Figura 10 são apresentadas as funções prontas, utilizadas na etapa de seleção de amostra, com seus respectivos pacotes e descrições:

Função	Pacote	Descrição
read_excel("<nome_do_arquivo>.xlsx")	readxl	Utilizada para carregar o arquivo no software R.
count(dados, “tipo”)	plyr	Utilizada para contar a quantidade de dados em cada categoria “tipo”.
difftime(var_1, var_2, units = "weeks")	base	Função utilizada para calcular em semanas a diferença entre duas datas (var_1 - var_2).
sapply(dados, function(x) sum(is.na(x)))	base	Exibe uma contagem de dados faltantes para cada variável.

‘

na.omit(dados)	stats	Utilizada para retirar linhas com dados faltantes do banco de dados.
pipe (%>%)	magrittr	Operador utilizado para encadear o resultado anterior na função seguinte.
mutate_at(fatores, list(~factor(.)))	dplyr	Edita colunas específicas para se tornarem um fator.

Figura 10 - Resumo das funções prontas, utilizadas na seleção da amostra

Na Figura 11 são apresentadas as funções montadas, utilizadas na etapa de seleção de amostra, com suas respectivas descrições:

Função	Descrição
dados=dados[,-c(which(names (dados) %in% c("cod_cli", "cep_com", "dt_admi", ”dtcad”, ”dtnasc”)))]	Na formatação atual, serve para retirar as variáveis descritas do banco de dados.
fatores = names(dados) [!names(dados) %in% c("idade_cad")] dados = dados %>% mutate_at (fatores, list(~factor(.)))	Na forma atual, serve para transformar as variáveis em fatores, exceto a variável descrita.

Figura 11 - Resumo das funções montadas, utilizadas na seleção da amostra

5.2 Análise Preliminar

Na Figura 12 são apresentadas as funções prontas, utilizadas na etapa de análise preliminar, com seus respectivos pacotes e descrições:

Função	Pacote	Descrição
plot_boxplot(dados,by="tipo")	dataexplorer	Utilizada para plotar um boxplot detalhando a variável tipo.
ggplot(dados, aes(y = sexo, fill = tipo))	ggplot2	Usada para declarar o banco de dados e o conjunto de estéticas do gráfico.
geom_bar(position = "fill")	ggplot2	Utilizado para plotar um histograma com a contagem dos indivíduos.
woebin(dados, y = "tipo", x = c("uf_natu"), check_cate_num = F)	scorecard	Utilizado para calcular o IV de variáveis com mais de cinco fatores.
select(idade_cad, y=tipo)	dplyr	Utilizada para selecionar variáveis.
floor(df$idade_cad)	base	Arredonda os valores em x para o número inteiro mais próximo, menor ou igual a ele.

‘

pivot_longer(cols = c(1), names_to = "Var", values_to = "Valor")	tidyr	Aumenta os dados, aumentando o número de linhas e diminuindo o número de colunas.
group_by(Var,Valor)	dplyr	Utilizada para agrupar um conjunto de dados existente por uma ou mais variáveis.
summarise(Total = n(), Bons = sum(y==1), Maus = Total - Bons)	dplyr	Utilizada para criar um novo conjunto de dados que contém um resumo estatístico dos dados de entrada.
full_join(dscTot, by = "Var")	dplyr	Todas as observações dos conjuntos são mantidas na junção, com correspondências incluídas, se houverem.
mutate(RR = (Bons/BonsTot)/(Maus/MausTot))	dplyr	Utilizada para criar novas colunas em um conjunto de dados que são funções de variáveis existentes.
unique(dsc$Var)	base	Utilizada para retornar um banco de dados sem linhas duplicadas.
length(unique(df[[i]]))	base	Utilizada para obter o número de elementos do objeto.
filter(Var == i)	dplyr	Usada para manter todas as linhas que satisfazem a condições indicada.
geom_line()	ggplot2	Traça uma linha em todos os pontos na ordem em que eles aparecem no eixo x
grid.arrange(grobs=plots)	gridExtra	Usada para organizar múltiplos gráficos em uma única grade.
seq(25,65,10)	base	Gera uma sequência de números do primeiro argumento ao segundo argumento com o intervalo do terceiro argumento.
cut(df1$idade_cad, kortes, include.lowest = T, left=T)	base	Usada para dividir um vetor numérico em intervalos

Figura 12 - Resumo das funções prontas, utilizadas na análise preliminar

Na Figura 13 são apresentadas as funções montadas, utilizadas na etapa de análise preliminar, com suas respectivas descrições:

‘

Função	Descrição
df = dados %>% select(idade_cad, y=tipo) df$idade_cad <- floor(df$idade_cad) dsc0 = df %>% pivot_longer(cols = c(1), names_to = "Var", values_to = "Valor") dsc <- data.frame() dsc = dsc0 %>% group_by(Var,Valor) %>% summarise(Total = n(), Bons = sum(y==1), Maus = Total - Bons) dscTot = dsc0 %>% group_by(Var) %>% summarise(TotalPop = n(), BonsTot = sum(y==1), MausTot = TotalPop - BonsTot) dsc = dsc %>% full_join(dscTot, by = "Var") dsc = dsc %>% mutate(RR = (Bons/BonsTot)/(Maus/MausTot)) dsc$RR[dsc$RR == Inf] = 4 vars = unique(dsc$Var) plots = list() for(i in vars){ k = length(unique(df[[i]])) k = floor(k/3) plots[[i]] = dsc %>% filter(Var == i) %>% ggplot(aes(x=Valor, y=RR)) + geom_line() + scale_y_continuous(limits = c(0,4)) + theme_bw() + labs(x = i, y="RR", title=i) } gridExtra::grid.arrange(grobs=plots)	Na forma atual serve para calcular e plotar o Risco Relativo para cada valor da variável idade_cad.
df1 = dados %>% dplyr::select(idade_cad, y=tipo) kortes = seq(25,65,10) kortes = c(0,kortes,100) nomeGrupo = cut(df1$idade_cad, kortes, include.lowest = T, left=T) df1 = df1 %>% mutate(Grupo = cut(df1$idade_cad, kortes, labels = F, include.lowest = T, left=T), nomeGrupo) df1 = df1 %>% group_by(Grupo) %>% summarise(Total = n(), Bons = sum(y==1), Maus = Total - Bons) %>% mutate(pBons = Bons/sum(Bons), pMaus = Maus/sum(Maus), RR = pBons/pMaus) df1 %>% ggplot(aes(x=Grupo, y=RR)) + geom_line() + theme_bw() + geom_point() + scale_x_continuous(breaks = seq(1,6), labels = unique(nomeGrupo)) + scale_y_continuous(breaks = seq(0,4,0.5), labels = seq(0,4,0.5)) +geom_hline(linetype=2, colour="red", yintercept = 1)	Na forma atual serve para agrupar os valores de idade em grupos e após calcular e plotar o Risco Relativo do novo agrupamento das idades.

Figura 13 - Resumo das funções montadas, utilizadas na análise preliminar

‘

5.3 Construção do Modelo

Na Figura 14 são apresentadas as funções prontas, utilizadas na etapa de construção do modelo, com seus respectivos pacotes e descrições:

Função	Pacote	Descrição
iv(dados, y = "tipo", order = T)	scorecard	Retorna o Valor da Informação de todas as variáveis presentes no banco de dados.
train_test_split(dados, prop = 0.7)	creditmodel	Utilizado para fazer a segmentalçao dos grupos de treino e teste do modelo.
modelo = glm(formula = tipo ~ ., data = ttreino, family = binomial (link = "logit"))	stats	Serve para ajustar o modelo de regressão logística.
check_collinearity(modelo)	performance	Serve para calcular o Fator de Inflação, indicador usado para analisar a multicolinearidade do modelo.
df = step(dados, scope = 'upper')	stats	Seleciona um modelo baseado em fórmula por AIC.

Figura 14 - Resumo das funções prontas, utilizadas na construção do Modelo

5.4 Escolha do Modelo

Na Figura 15 são apresentadas as funções prontas, utilizadas na etapa de seleção de amostra, com seus respectivos pacotes e descrições:

Função	Pacote	Descrição
predict(modelo,type= 'response',teste)	stats	Função para testar o modelo.
prediction(pred, truth)	rocr	Usado para avaliar a performance de modelos de classificação
performance(predob, "tpr", "fpr")	rocr	Usado para avaliar a performance de modelos de classificação
auc(truth, pred)	proc	Usado para calcular a área sob a curva ROC
rocplot(score, tipo, main="ROC")	rocr	Usado para plotar a curva ROC
geom_density(alpha=0.5)	ggplot2	Utilizado para plotar um gráfico de densidade da pontuação pela variável resposta.

‘

quantile(teste$yChapeu, seq(0,1,1/10))	stats	Produz quantis de amostra correspondentes às probabilidades fornecidas
cumsum(txMau)	base	Retorna um vetor cujos elementos são as somas cumulativas dos elementos do argumento

Figura 15 - Resumo das funções prontas, utilizadas na escolha do modelo

Na Figura 16 são apresentadas as funções montadas, utilizadas na etapa de escolha do modelo, com suas respectivas descrições:

Função	Descrição
rocplot <- function(pred, truth, ...) { predob = prediction(pred, truth) perf = ROCR::performance(predob, "tpr", "fpr") plot(perf, ...) area <- auc(truth, pred) area <- format(round(area, 4), nsmall = 4) text(x=0.8, y=0.1, labels = paste("AUC =", area)) segments(x0=0, y0=0, x1=1, y1=1, col="gray", lty=2)} rocplot(teste$score, teste$tipo, col="blue",main="ROC-Teste")	Na forma atual serve para calcular teste ROC e plotar o gráfico AUC.
teste$predito<-ifelse(teste$score>=0.5,1,0) tab_teste<-table(teste$tipo,teste$predito) taxaacerto_teste=(tab_teste[2,2]+tab_teste[1,1])/sum(tab_teste) print(paste0("Acurácia: ", round(taxaacerto_teste,2)))	Na forma atual retorna os valores previstos e observados do modelo, apresentando a acurácia do mesmo.
kortes = quantile(teste$yChapeu, seq(0,1,1/10)) teste = teste %>% mutate(Grupo = cut(yChapeu,breaks =kortes, left=T, labels=F, include.lowest=T)) dsc = teste %>% group_by(Grupo) %>% summarise(Tot = n(), Bons = sum(tipo == 1), Maus = sum(tipo==0)) %>% mutate(pMau = Maus/Tot) dsc %>% ggplot(aes(x=Grupo, y=pMau)) + geom_line()+geom_point()+theme_bw() +labs(x="Grupo", y="Proporção de Inadimplentes")	Na forma atual, serve para dividir a pontuação em decis, calcular a taxa de inadimplência nos mesmos e plotar o gráfico do resultado.
dsc = dsc %>% mutate(txMau = Maus/sum(Maus), txBom = Bons/sum(Bons), acumMau = cumsum(txMau), acumBom = cumsum(txBom), KS = abs(acumBom - acumMau)) dfPlt = pivot_longer(dsc, cols = c(8,9), names_to = 'Tipo', values_to = 'Valores') dfPlt %>% ggplot(aes(x=Grupo, y=Valores, colour = Tipo)) + geom_line() + annotate(geom = 'label',x y = 0.5,label = paste0("KS = ", round(max(dsc$KS),2)))	Na forma atual serve para calcular o teste de Kolmogorov-Smirnov (KS) para duas amostras e plotar o gráfico do mesmo.

Figura 16 - Resumo das funções montadas, utilizadas na escolha

‘

6. Considerações Finais

A modelagem de crédito é cada vez mais importante para bancos e instituições financeiras mitigarem suas perdas durante a liberação de crédito, aumentando o lucro e permitindo oferecer condições melhores aos bons pagadores. O objetivo desse estudo foi apresentar pacotes e funções do *software R* para construir modelos de pontuação de risco de crédito, apresentando possibilidades para vencer as etapas desde a concepção do objetivo e a delimitação da população até a implementação do modelo escolhido junto a empresa.

Nesse trabalho, foi realizado um exemplo de construção do modelo, em que os resultados foram de 66% de acurácia e 30% no teste KS, o que mostram bons resultados para o banco de dados analisado. Porém o maior resultado foi a grande quantidade de possibilidades de funções e pacotes que podem ser utilizadas para resolver o processo de modelagem. Com isso, pode-se montar no capítulo anterior um material de apoio apresentando e descrevendo algumas dessas possibilidades.

Como sugestões para trabalhos futuros são sugeridos: (1) a utilização de outros pacotes e funções, novos ou que não tenham sido utilizados nesse trabalho, como o pacote H2O, para realizar o processo de modelagem, sendo possível comparar qual a melhor função para um mesmo caso; (2) o uso de outras técnicas de modelagem, além da regressão logística, para realizar o modelo de pontuação de crédito, incrementando no material de apoio para auxiliar ainda mais cientistas de dados a realizarem análises como essas; (3) desenvolver as funções montadas neste trabalho no formato de funções do R, para difundir novas funções; (4) apresentar funções e pacotes para trabalhar com bancos de dados com um grande número de variáveis, buscando facilitar a avaliação da estabilidade dos dados ao longo do tempo ou a seleção automática de variáveis com base no seu IV, por exemplo; (5) apresentar pacotes e funções para trabalhar com variáveis desbalanceadas e variáveis que contenham *missings*.

‘

7. Referências

BACCONI, G. E. e GOUVÊA, M. A.; **Análise de Risco de Crédito com o Uso de Modelos de Regressão Logística e Redes Neurais**. São Paulo: FEA/USP, 2005.

BRITO, G. A. S. ASSAF NETO; **Modelo de classificação de risco de crédito de empresas**. Revista Contabilidade & Finanças [online]. 2008, v. 19, n. 46 [Acessado 20 setembro 2022], pp. 18-29. Disponível em: <https://doi.org/10.1590/S1519-70772008000100003>.

BRITO, G. A. S. ASSAF NETO, A. e CORRAR, L. J.; **Sistema de classificação de risco de crédito: uma aplicação a companhias abertas no Brasil**. Revista Contabilidade & Finanças [online]. 2009, v. 20, n. 51 [Acessado 20 setembro 2022], pp. 28-43. Disponível em: <https://doi.org/10.1590/S1519-70772009000300003>.

CORRAR, L. J.; PAULO, E.; DIAS FILHO, J. M. **Análise Multivariada: para cursos de Administração**, Ciências Contábeis e Economia. São Paulo: Atlas, 2007.

DUARTE JUNIOR, A. M.; **Risco: definições, tipos, medição e recomendações para o seu gerenciamento**. Resenha BM&F, n.144, 1996. Disponível em: <https://pt.scribd.com/document/85710096/Risco-Definicoes-tipos-medicoes-e-recomendacoes-para-o-seu-Gerenciamento >. Acesso em: 20/08/2022.

GERHARDT, T. E.; SILVEIRA, D. T.; **Planejamento e Gestão para o Desenvolvimento Rural da SEAD/UFRGS**. Porto Alegre: Editora da UFRGS, 2009.

GOUVÊA, M. A.; GONÇALVES, E. B. **Análise de Risco de Crédito com o Uso de Modelos de Redes Neurais e Algoritmos Genéticos**. In: IX SEMEAD – Seminários em Administração FEA-USP, 2006, São Paulo. Anais.

KAUFFMANN, L. H. O. **Uma abordagem Forward-Looking para estimar a PD segundo IFRS9**. 2017. Dissertação (Mestrado em Matemática, Estatística e Computação) - Instituto de Ciências Matemáticas e de Computação, University of São Paulo, São Carlos, 2017. doi:10.11606/D.55.2018.tde-06112018-182558. Acesso em: 2023-02-21.

LEWIS, E. M. **An Introduction to Credit Scoring**. San Rafael: Fair, Isaac and Co., Inc. 1992.

‘

LÜDECKE et al., (2021). performance: **An R Package for Assessment, Comparison and Testing of Statistical Models**. Journal of Open Source Software, 6(60), 3139. https://doi.org/10.21105/joss.03139

MAYER, F.; ZEVIANI, W. Pesquisa **reproduzível com o R: de documentos dinâmicos a pacotes**. LEG/UFPR. 2016. Disponível em < http://cursos.leg.ufpr.br/prr/capPacR.html >. Acesso em: 21/08/2022.

NEVES JUNIOR, J.; MATSUMOTO, A. S.; ARAUJO, S. S. R. **Grau de investimento: uma análise comparativa do desempenho Brasil em relação à Argentina, Chile, México, Uruguai e Venezuela, na visão do rating da Moody's Group**. In: CONGRESSO INTERNACIONAL DE COSTOS, 10, 13, 14 e 15 de junho de 2007, Lyon/França. Anais... Lyon/França: Instituto Internacional de Custos, 2007.

OWEN, W. J.; **The R Guide**. Department of Mathematics and Computer Science - University of Richmond. 2010. Disponínel em: < https://cran.r-project.org/doc/contrib/Owen-TheRGuide.pdf >. Acessado em 20/09/2022.

PEREIRA, A. P. F.; BARROSO, M. H.; NEPOMUCENO FILHO, F. **Uso do Credit Score na Análise de Crédito de Pessoa Física**. In: Congresso Nacional de Excelência em Gestão, nov. 2002, Niterói, RJ.

PICININI, R.; OLIVEIRA, G. M. B.; MONTEIRO, L. H. A. **Mineração de Critério de Credit Scoring Utilizando Algoritmos Genéticos**. In: VI Simpósio Brasileiro de Automação Inteligente, 2003, Bauru.

POLO TCF, M. H.; **Aplicações da curva ROC em estudos clínicos e experimentais**. J Vasc Bras. 2020;19: e20200186. https://doi.org/10.1590/1677-5449.200186.

QUEIROZ, R. S. B. **A importância dos modelos de Credit Scoring na concessão de crédito ao consumidor no varejo**. In: IX SEMEAD – Seminários em Administração FEA-USP, 2006, São Paulo.

R CORE TEAM. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria, 2011. URL: http://www.R-project.org/.

‘

SCHRICKEL, W. K. **Análise de Crédito: concessão e gerência de empréstimos**. 3.ed. São Paulo: Atlas, 1997.

SELAU, L. P. R. **Construção de modelos de previsão de risco de crédito**. Porto Alegre: UFRGS, 2008. Dissertação (Mestrado em Engenharia da Produção), Programa de Pós-Graduação em Engenharia de Produção, Universidade Federal do Rio Grande do Sul, 2008.

SHARMA, D.; **Guide to Credit Scoring in R**. R Core Team. 2009. Disponível em: < https://cran.r-project.org/doc/contrib/Sharma-CreditScoring.pdf >. Acessado em 20/09/2022.

SZEPANNEK, G; **An Overview on the Landscape of R Packages for Credit Scoring**. Volume 2. The R Journal, 2020.

THOMAS, C. L. **A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers**. International Journal of Forecasting, v.16, n.2, p.149-172, 2000.

VASCONCELLOS, R. S. **Modelos de Escoragem de Crédito Aplicados a Empréstimo Pessoal com Cheque**. Rio de Janeiro: FGV, 2004. Dissertação (Mestrado em Finanças e Economia Empresarial), Escola de Pós-Graduação em Economia, Fundação Getúlio Vargas, 2004.