Universidade de São Paulo
Instituto de Matemática e Estatística
Bacharelado em Matemática Aplicada e Computacional

Modelagem Preditiva e Avaliação de Risco de Crédito em Populações de Alta Probabilidade de Inadimplência

Gustavo Soares Gomes

Monografia Final

MAP 2040 — Trabalho de Formatura Supervisionado

Supervisor: Prof. Dr. Antonio Geraldo da Rocha Vidal
Cossupervisor: Prof. Dr. Luís Gustavo Esteves

São Paulo
2024

"O dinheiro é um bom criado, mas um mau senhor."
– Francis Bacon

Agradecimentos

O sonho do careta é a realidade do maluco.
— Raul Seixas

Agradeço, primeiramente, a Deus pelo dom da vida e pela saúde, sem os quais nada seria possível. Sou imensamente grato aos meus pais, pelo esforço, dedicação e amor durante minha criação, ao meu irmão, pelo apoio incondicional, à minha avó, pelo carinho e pela presença constante em minha vida e à minha companheira pelo apoio durante este ano.

Expresso minha gratidão aos meus colegas de curso, especialmente João Pedro e José Luiz, pelo apoio e companheirismo ao longo dessa jornada acadêmica. Agradeço também aos docentes Luis Gustavo Esteves, Leonardo T. Rolla e Airlane Pereira Alencar, pela confiança e pela oportunidade de colaborar como monitor em suas disciplinas, experiência que contribuiu significativamente para o meu crescimento acadêmico e pessoal.

De forma especial, reconheço a solicitude dos professores Luis Gustavo Esteves e Antonio Geraldo da Rocha Vidal, cujas orientações foram essenciais para esclarecer minhas dúvidas e aprimorar este projeto. A todos que, direta ou indiretamente, contribuíram para esta conquista, deixo meu mais sincero obrigado.

Resumo

Gustavo Soares Gomes. **Modelagem Preditiva e Avaliação de Risco de Crédito em Populações de Alta Probabilidade de Inadimplência**. Monografia (Bacharelado). Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2024.

Este trabalho tem como objetivo a implementação e análise de técnicas para avaliação e concessão de crédito para públicos com alto risco de inadimplência, visando comparar a eficácia de diferentes modelos preditivos. A base de dados utilizada foi a do Lending Club obtida através da plataforma Kaggle, que representa a população em estudo, e foram desenvolvidas técnicas de análise com foco no aprimoramento dos modelos tradicionais, como a Regressão Logística e a Floresta Aleatória, amplamente utilizados no mercado financeiro. Para a avaliação da performance dos modelos, foi realizada uma adaptação de indicadores de acurácia, ajustados ao contexto específico deste estudo. Os resultados indicam que a Floresta Aleatória permanece como o modelo mais eficaz para a análise de risco de crédito em populações de alto risco de inadimplência, além disso o modelo proposto superou a Regressão Logística, apresentando um desempenho superior na previsão de risco de crédito, o que sugere que técnicas mais avançadas, como a Floresta Aleatória, são mais apropriadas para cenários de alto risco de inadimplência, oferecendo maior precisão nas decisões de concessão de crédito.

Palavras-chave: crédito. risco de crédito. análise de crédito. regressão logística. floresta aleatória.

Abstract

Gustavo Soares Gomes. . Capstone Project Report (Bachelor). Institute of Mathematics and Statistics, University of São Paulo, São Paulo, 2024.

This study aims to implement and analyze techniques for credit assessment and granting for high-risk default populations, comparing the effectiveness of different predictive models. The dataset used was from Lending Club, obtained through the Kaggle platform, representing the studied population. Analysis techniques were developed focusing on improving traditional models such as Logistic Regression and Random Forest, which are widely used in the financial market.

To evaluate the performance of the models, accuracy indicators were adapted and tailored to the specific context of this study. The results indicate that Random Forest remains the most effective model for credit risk analysis in high-risk default populations. However, the proposed model outperformed Logistic Regression, demonstrating superior performance in credit risk prediction. The conclusions suggest that more advanced techniques, such as Random Forest, are better suited for high-risk default scenarios, offering greater precision in credit granting decisions.

Keywords: Credit. credit risk. credit analysis. logistic regression. random forest.

Lista de abreviaturas

BACEN	Banco Central do Brasil (*Central Bank of Brazil*)
PIB	Produto Interno Bruto (*Gross Domestic Product*)
PF	Pessoa Física (*Natural Person*)
CR	Comprometimento de Renda (*Income Commitment*)
EV	Endividamento Familiar (*Household Debt*)
LGPD	Lei Geral de Proteção de Dados (*General Data Protection Law*)
IME	Instituto de Matemática e Estatística
USP	Universidade de São Paulo

Lista de símbolos

α	Parâmetro de Flexibilidade

Lista de figuras

Lista de tabelas

Lista de programas

Sumário

Introdução

De acordo com dados divulgados pelo Banco Central do Brasil (BACEN) e reportados pela CNN Brasil, as concessões de empréstimos apresentaram um aumento de 3,7% em julho de 2024, quando comparado ao mês anterior. Esse crescimento evidencia a importância do mercado de crédito no Brasil, que não só impulsiona a economia, mas também desempenha um papel fundamental no consumo das famílias e nos investimentos empresariais. Contudo, para entender plenamente a dinâmica do crédito, é essencial analisar de forma aprofundada os processos de decisão adotados pelas instituições financeiras, especialmente considerando as mudanças significativas no perfil dos tomadores de empréstimos no Brasil nos últimos anos.

No atual cenário macroeconômico brasileiro, as instabilidades causadas por fatores como desemprego, déficit fiscal e dívida pública acarretam aumento da inflação e a elevação das taxas de juros, o que têm elevado o custo do crédito. Em consequência, houve uma siginificativa piora do perfil de risco de crédito, já que devido às essas instabilidades registrou-se um aumento dos índices de inadimplência e endividamento familiar. Esse cenário representa um desafio tanto para as instituições financeiras quanto para o desenvolvimento de políticas de crédito que favoreçam a inclusão financeira de maneira sustentável.

Diante desse contexto, o presente trabalho busca responder à seguinte questão: Dada a mudança no perfil médio dos tomadores de empréstimos no Brasil, será necessário que as instituições financeiras adotem modelos mais conservadores na concessão de crédito?

Para abordar essa questão, o Capítulo 1 apresenta uma revisão sobre o conceito de crédito e uma análise da conjuntura econômica atual do Brasil, com foco nos fatores que impactam a inadimplência. O Capítulo 2 detalha as principais técnicas de análise de crédito utilizadas atualmente no mercado, como a Regressão Logística e a Floresta de Decisão, com destaque para o rigor matemático necessário para a implementação dessas metodologias.

No Capítulo 3, propõe-se utilizar uma base de dados extraída da instituição financeira norte-americana Lending Club, como uma proxy para o perfil da população tomadora de empréstimos no Brasil. Essa base de dados, originada no contexto da crise financeira de 2008 nos Estados Unidos, será tratada ao longo do capítulo, resultando em duas bases (uma de treino e uma de teste), cada uma com 10 variáveis, que servirão como entrada para as técnicas de análise de crédito.

O Capítulo 4 descreve o passo a passo do desenvolvimento de um modelo de risco conservador, utilizando os conhecimentos estatísticos adquiridos ao longo da minha formação. Esse modelo busca responder aos desafios impostos pelas condições econômicas atuais,

priorizando a segurança nas decisões de concessão de crédito.

Por fim, o Capítulo 5 apresenta as métricas de eficiência que serão utilizadas para avaliar o desempenho dos modelos desenvolvidos. Neste capítulo, será implementada tanto a técnica de risco conservador proposta no Capítulo 4 quanto as técnicas tradicionais mencionadas no Capítulo 2, com o objetivo de comparar sua eficácia na previsão da inadimplência e no controle do risco de crédito.

Capítulo 1

Crédito

1.1 Conceito de Crédito

A palavra “crédito” sempre foi de uso frequente em operações no mercado, entretanto são poucos os que conhecem o real significado e as nuances desse tipo de operação. De acordo com Júnior, 2014, o real desenvolvimento de concessões de crédito ocorreu durante a Revolução Industrial devido ao fato de os donos de indústrias recorrerem a grandes detentores de capital a fim de financiar seus maquinários para aumentar a produtividade, pois a receita gerada pela indústria não era suficiente para fazer frente aos investimentos necessários para aquisição do novo maquinário.

Ainda dentro desse contexto, o autor afirma que o crédito nada mais é do que a troca de uma prestação atual por uma prestação futura, onde a diferença entre esses dois instantes possui um custo, que no caso das instituições financeiras e bancos é chamado de “juros": a margem de lucro da instituição decorrente da concessão do empréstimo.

1.2 Tipos de Crédito

De acordo com Gastaldi, 2005 existem três formas de classificação do crédito. A primeira diz respeito ao tipo do devedor, que pode ser uma instituição pública (União, Estado), uma instituição privada (bancos, empresas, startups) ou ainda um particular (pessoa física). A segunda diz respeito ao uso a que se destina o empréstimo, podendo ser: consumitivo (quando a quantia do empréstimo é destinada a bens de consumo e que não necessariamente produzirão algum capital futuro) ou produtivo (quando o capital é destinado a bens e/ou processos cuja finalidade é gerar mais capital, um exemplo é quando uma pequena empresa realiza um crédito a fim de expandir suas operações).

Ainda há uma terceira classificação que diz respeito à garantia que é fornecida pelo solicitante do empréstimo. Dizemos que o crédito é pessoal quando sua fundamentação é baseada unicamente na palavra do tomador (garantia de natureza abstrata), enquanto que no crédito real o tomador vincula alguma garantia (que pode ser um imóvel ou automóvel, por exemplo) que lhe pode ser tomada em caso de inadimplência da obrigação. Justamente

pela falta de uma garantia, é natural no mercado que o crédito pessoal tenha taxas de juros maiores do que o crédito real.

1.3 Cenário do Crédito no Brasil

De acordo Carvalho *et al.*, 2007, o crédito é de grande importância em uma economia capitalista, pois atua como elemento que gera dinamismo no mercado. No contexto brasileiro observa-se a importância dos vários tipos de crédito em diversas áreas da economia, dentre as quais se destacam: agronegócio, indústria, comércio, micro e pequenas empresas, infraestrutura e habitação.

Dentro desse contexto, de acordo com o BACEN, o saldo de crédito é definido como o saldo em final de período das operações de crédito contratadas no Sistema Financeiro Nacional. No gráfico a seguir, podemos notar um aumento do saldo de crédito às pessoas físicas (PF) em relação ao PIB, o que evidencia um maior endividamento das famílias, por isso vemos que essa modalidade de crédito têm crescido bastante nos últimos anos e já está em um patamar representativo do PIB brasileiro.

Dentro do crédito à pessoa física (PF) as modalidades mais representativas no Brasil são as que não envolvem garantia, como: crédito pessoal, cartão de crédito e cheque especial. Esses tipos específicos de empréstimos são caracterizados por taxas de juros muito altas e consequentemente aumento do comprometimento da renda da pessoa tomadora de crédito, o que em muitas vezes acaba levando a um ciclo vicioso de inadimplência.

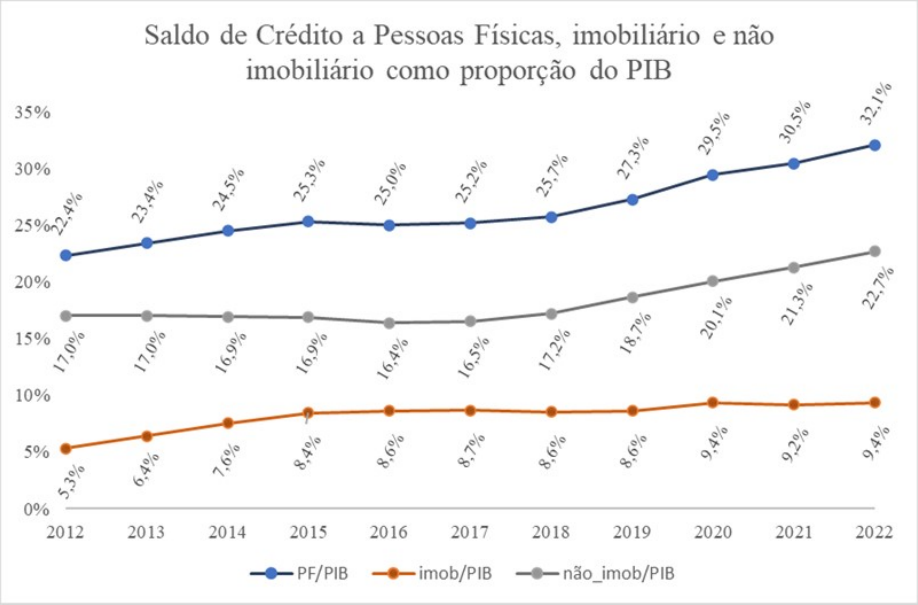
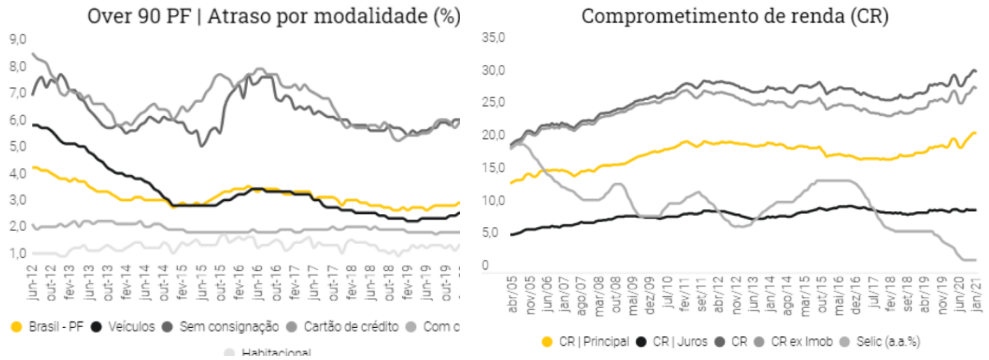


Figura 1.1: *Saldo de Crédito a PF segundo tipo de crédito. Fonte: BACEN*

Não à toa os índices de inadimplência de PF, Comprometimento de Renda (CR) e Endividamento Familiar (EV) vêm crescendo aceleradamente no Brasil últimos anos, como mostram as visões a seguir extraídas do BACEN. Tendo em vista o maior acesso ao crédito da população anteriormente desbancarizada, é de fundamental importância o desenvolvimento de técnicas de tomada de decisão de crédito para esses públicos com

maior inadimplência, pois a sua exclusão do mercado causaria um grande impacto negativo na cadeia de produção devido à desaceleração do consumo, ao mesmo tempo que deve-se tomar cuidado na concessão de crédito para clientes com perfil de alto risco, o que pode acarretar em diversos prejuízos, tais como: desaceleração da economia, aumento geral na taxa de juros e na falência de instituições bancárias/financeiras.



(a) *% contratos c/ atraso maior que 90 dias.* **(b)** *Comprometimento de Renda.*

Figura 1.2: *Over 90 PF e CR entre 2012 e 2019. Fonte: Expert XP com dados extraídos do BACEN*

Dentro desse contexto do aumento da inadimplência no Brasil, algumas estratégias têm sido tomadas tanto pelo Estado como também por parte das instituições privadas para tentativa de maior controle desses números. A primeira que podemos citar são os programas de educação financeira que auxiliam os consumidores na gestão de suas finanças e controle do seu endividamento, como exemplo temos o Programa de Educação Financeira do Banco Central (PEF-BC) que envolve ações que tem por objetivo a orientação da sociedade sobre assuntos financeiros, destacando o papel do BACEN como agente promotor da estabilidade econômica.

Embora esses programas sejam de grande validade, possuem resultados geralmente perceptíveis apenas no longo prazo e não conseguem ajudar instantaneamente as pessoas endividadas. Uma das alternativas então é a renegociação de dívidas, artifício no qual muitas instituições financeiras em conjunto com o governo federal ou com programas próprios de renogiciação de dívidas, oferecem opções mais acessíveis para os credores quitarem seus débitos, como: maior parcelamento da dívida ou até perdão de parte dela.

Um instrumento que está tanto relacionado a educação financeira como também ao resultado imediato da concessão de um empréstimo é a análise de crédito. Quando clientes com perfil mais arriscado recebem scores baixos de birôs, existe a reflexão sobre sua situação financeira. Por outro lado, a análise de crédito é fundamental para discriminar os clientes para as empresas, no sentido de ajudá-las a decidir se vale a pena (considerando fatores de risco e rentabilidade do negócio) assumir o risco de conceder-lhe um empréstimo ou não e em caso positivo, estipular qual a taxa que remunera o risco de inadimplência deste cliente. Por esse motivo, é essencial que estudemos os métodos empregados atualmente na análise de concessão de crédito, assim como tentemos propor possíveis melhorias.

Capítulo 2

Análise de Crédito

2.1 Riscos da Concessão de Crédito

Como comentado anteriormente, a concessão de crédito traz naturalmente uma disposição ao risco por envolver a expectativa de retorno do patrimônio, portanto é necessário elaboração e aprimoramento de métodos de administração desse risco por parte das instituições concedentes dos empréstimos.

Os subtipos desse risco podem ser classificados segundo Figueiredo, 2001 como:

- **Risco de inadimplência:** risco do não pagamento por parte do tomador em uma operação de crédito - empréstimo, financiamento, adiantamentos, operações de *leasing* - ou ainda a possibilidade de uma contraparte de um contrato ou emissor de um título não honrar seu crédito;
- **Risco de degradação de garantia:** risco de perdas em função das garantias oferecidas por um tomador deixarem de cobrir o valor de suas obrigações junto à instituição, em função de desvalorização do bem no mercado e dilapidação do patrimônio empenhado pelo tomador;
- **Risco de concentração de crédito:** possibilidade de perdas em função da concentração de empréstimos e financiamentos em poucos setores da economia, classes de ativos ou de empréstimos elevados para um único cliente ou grupo econômico;
- **Risco de degradação do crédito:** perda pela queda na qualidade creditícia do tomador de crédito, emissor de um título ou contraparte de uma transação, ocasionando uma diminuição no valor de suas obrigações. Esse risco pode acontecer em uma transação do tipo de aquisição de ações ou de títulos soberanos que podem perder valor.
- **Risco soberano**: risco de perdas em transações internacionais quando um país não consegue ou não está disposto a cumprir suas obrigações financeiras, como o pagamento de dívidas externas, títulos ou empréstimos obtidos no mercado internacional.

Dentro do contexto de crédito pessoal sem garantia, iremos trabalhar no controle do risco de inadimplência, utilizando e ressignificando técnicas de Regressão Logística e *Random Forest* para identificar clientes mais propensos a honrar seus créditos durante uma operação.

Uma análise qualitativa da concessão do empréstimo pode ser realizada por meio de analistas a partir do julgamento de quão propensa ao risco é uma operação. Esta abordagem por um lado humaniza o julgamento e gera uma visão mais ampla do cliente, porém por outro pode ser influenciada por viéses dos analistas e possui alto custo de manutenção.

Em uma outra perspectativa temos a análise quantitativa, que geralmente é feita utilizando-se modelos de avaliação do cliente para prever a possibilidade de não pagamento do crédito a partir de variáveis como renda, endividamento, entre outras. Dentre os modelos mais utilizados no mercado estão o Credit Scoring e a Análise Discriminante.

De acordo com Gonçalves *et al.*, 2013 geralmente as instituições financeiras utilizam as duas análises combinadas, sendo a quantitativa utilizada em massa e a qualitativa utilizada em casos de exceção, como por exemplo clientes mais arriscados ou clientes políticamente expostos. Neste trabalho, iremos focar nos métodos quantitativos, porém é fundamental reconhecer a importância do método qualitativo para prevenção de fraudes e correção de possíveis erros nos modelos analíticos.

2.2 Modelos de Credit Scoring

Devido a alta quantidade de tomadores de crédito, as instiuições financeiras geralmente possuem sistemas automatizados que atribuem uma nota para cada tomador a fim de avaliar se é compensatório ou não assumir o risco de perda daquela transação. Esta classificação ocorre pelos modelos de Credit Scoring que podem ser:

- **Simples:** O resultado do modelo atribui a cada tomador o escore 0 (relativo a clientes com alta probabilidade de inadimplência) ou escore 1 (relativo a clientes com baixa probabilidade de inadimplência);
- **Complexo:** O resultado do modelo atribui a cada possível tomador um escore que varia dentro de um intervalo de valores (geralmente de 0 a 100) e uma segmentação é feita tomando por principio que clientes com maiores escores tendem a ter uma perda por inadimplência menor comparados aos clientes com baixos escores.

No contexto brasileiro, o processo de estabilização econômica com o Plano Real (1994) proporcionou um ambiente mais favorável para a concessão de crédito o que abriu espaço para importação de tecnologias de Credit Scoring dos EUA e da Europa.

De acordo com Gonçalves *et al.*, 2013 existem sete etapas principais para a aplicação de um modelo de Credit Scoring em uma instituição:

1. **Levantamento de base histórica de clientes:** A suposição fundamental em um modelo de avaliação de crédito é que os clientes tenham o mesmo comportamento ao longo do tempo. Em nosso caso, iremos considerar clientes de uma mesma empresa em um período de tempo pré-fixado, portanto é esperado que tenham o mesmo comportamento de exposição a inadimplência ao longo dos anos;

2. **Classificação dos clientes de acordo com o padrão de comportamento e a definição da variável resposta:** As instituições possuem sua própria política de crédito e os conceitos de bons e maus pagadores podem variar. Dentro desse trabalho, estaremos adotando o conceito de "bom" pagador àquele que concluiu o pagamento do seu empréstimo independentemente de atrasos em algumas parcelas e o "mau" pagador será definido como aquele em que não se espera mais recuperar a totalidade da quantia emprestada com acréscimo de juros;
3. **Seleção de amostra aleatória representativa da base histórica:** É importante que as amostras de "bons" e "maus" tomadores possuam o mesmo tamanho a fim de se evitar viéses nas premissas dos modelos. Em nossa análise estaremos utilizando uma quantidade igual de 175 mil bons e maus pagadores para o treino dos modelos;
4. **Análise descritiva e preparação dos dados:** Consiste em analisar cada variável a ser utilizada no modelo segundo critérios estatísticos. Neste projeto uma robusta análise descritiva foi feita com os dados no capítulo 3 a fim de encontrar as variáveis mais relevantes para compor os modelos desenvolvidos;
5. **Escolha e aplicação das técnicas a serem utilizadas para a construção do modelo:** Nesse estudo utilizaremos as técnicas de Regressão Logística, Floresta de Decisão e uma técnica construida pelo próprio autor;
6. **Definição dos critérios de comparação dos modelos:** Utilizaremos alguns indicadores de acertos baseados na matriz de confusão. É importante notar que dentro do nosso contexto, existe um custo maior em se classificar erroneamente um mau pagador do que um bom pagador, de maneira que devemos olhar com uma perspectiva apropriada para essas métricas de acurácia;
7. **Seleção e Implantação do melhor modelo:** De acordo com critérios definidos, o melhor modelo é escolhido; para implantá-lo, o usuário deve possuir sistemas que possam comportar o algoritmo final.

A fim de ter uma visão mais ampla dos métodos matemáticos que permeiam as técnicas que serão usadas em nossos modelos, as próximas duas sessões se dedicam a explicar sucintamente a lógica por trás das técnicas de Regressão Logística e Árvores de Decisão.

2.3 Regressão Logística

2.3.1 Definição

No contexto em que queremos estimar o valor de uma variável Y como função de variáveis independentes $x_1, x_2, ..., x_k$ o método mais utilizado é a regressão linear, que tenta estimar o valor médio de Y condicionado às variáveis independentes como uma combinação linear destas:

$$\mathbb{E}(y|x_1, x_2, ...x_k, \beta) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_k$$

Entretanto, quando a variável Y pode assumir apenas valores dentro do conjunto $\{0, 1\}$, a regressão linear não se torna um modelo satisfatório para ajuste dos dados, já que olhando

sob uma perspectiva da plotagem dos dados sobre um gráfico de coordenadas cartesianas, os pontos estarão concentrados nos eixos $y = 0$ e $y = 1$. O mais recomendável então será considerar um modelo onde a decisão do valor esperado de Y depende da sua probabilidade e não diretamente das variáveis independentes:

$$\mathbb{E}(y|X,\beta) = \begin{cases} 1, & \text{se } \mathbb{P}(y|X,\beta) \geq 0.5 \\ 0, & \text{se } c.c \end{cases}$$

Entretanto essa probabilidade ainda é função do vetor X e dado o formato em que se distribuem os dados é natural pensar em formato de S ou de uma função de probabilidade acumulada para a função.

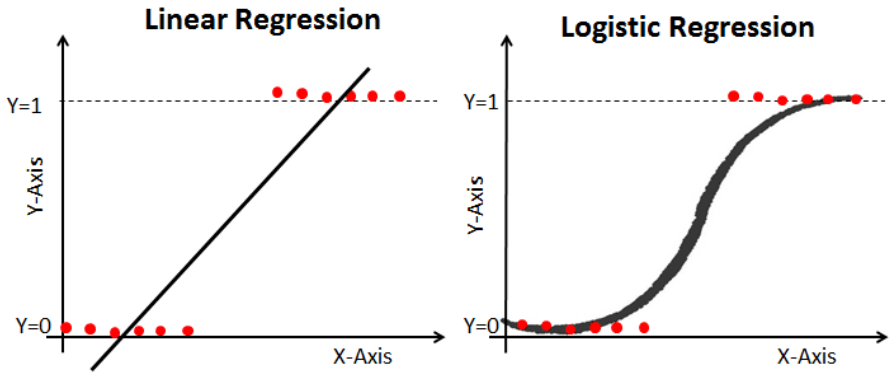


Figura 2.1: *Comparação entre regressão tradicional e a regressão logística. Fonte: Analytics Vidhya*

D.R. e E.J., 1989 discutem as várias possíveis funções que poderiam ser tomadas aqui, porém por questão de praticidade utilizaremos a curva logística, então:

$$\mathbb{P}(Y|x_1, x_2, ..., x_k, \beta) = \frac{1}{1 + e^{-g(x_1,x_2,...,x_k,\beta)}}$$

onde $g(x_1, x_2, ..., x_k, \beta) = \beta_0 + \beta_1 x_1 + ... + \beta_n x_k$.

2.3.2 Cálculo dos coeficientes β_i

Suponha que temos uma amostra $\{(X_i, y_i)\}_{i=1}^n$ de observações, onde a variável y possui valores 0 ou 1 e o vetor $X_i \in \mathbb{R}^k$ contempla os valores das k variáveis relativas ao indivíduo i. Iremos estimar os valores de β através da função dada a seguir pelo método de máxima verossimilhança.

$$l(\beta) = \prod_{i=1}^{n} P(Y = 1)^{y_i}[1 - P(Y = 1)]^{1-y_i}$$

Os β_0, β_1 que maximizam esta função são tais que também maximizam seu logaritimo natural, pois o logaritmo é uma função estritamente crescente. A condição de primeira

ordem em relação a β_j é que a primeira derivada parcial da função seja nula. (note que $P(Y = 1)$ depende dos β_i):

$$ln(l(\beta)) = \sum_{i=1}^{n} y_i[P(Y = 1)] + (1 - y_i)[1 - P(Y = 1)]$$

$$\begin{cases} \frac{\partial ln(l(\beta))}{\partial \beta_j} = \sum_{i=1}^{n} x_j[y_i - P(Y = 1)] = 0, \text{para } j = 1, 2, ..., n \\ \sum_{i=1}^{n} [y_i - P(Y = 1)] = 0, \text{para } j = 0 \end{cases}$$

Na regressão linear, essas equações são resolvidas de forma linear nos β_i, o que não ocorre aqui na regressão logística. Neste caso, o software calcula uma solução aproximada para os coeficientes β_i populacionais utilizando algorítimos de analise numérica, em especial o algoritmo de Newton-Raphson, cuja descrição mais detalhada se encontra em Geeks For Geeks, 2024.

2.3.3 Interpretação dos coeficientes β_i

Na regressão linear tradicional a interpretação do coeficiente β_i associado à variável x_i é que a variação em uma unidade na variável x_j resulta na variação de β_j unidades na variável Y. No caso da regressão logística, a interpretação dessa variação não é tão imediata pelo fato de estarmos lidando com probabilidades e por Y não depender linearmente da variável x_i.

Para estimarmos o efeito dessa variação vamos chamar $p = P(Y = y)$ e supor que aplicamos uma regressão logística para estimar o efeito dos regressores $x_1, x_2, ..., x_k$ na variável Y. Da definiçao de p, podemos escrever:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_k$$

Chamemos de p_n a nova distribuição de probabilidade decorrente da variação de 1 unidade na variável x_i:

$$\log\left(\frac{p_n}{1-p_n}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i + \beta_n x_k + \beta_i$$

$$\log\left(\frac{p_n}{1-p_n}\right) = \log\left(\frac{p}{1-p}\right) + \beta_i$$

$$\left(\frac{p_n}{1-p_n}\right)\left(\frac{1-p}{p}\right) = e^{\beta_i}$$

$$\frac{p_n}{p} = \left(\frac{1-p_n}{1-p}\right) e^{\beta_i}$$

Vemos então que a razão que compara as probabilidades antes e depois da variação em 1 unidade na variável x_i possui seu valor determinado também por p, isto é, a variação

impacta a curva de probabilidade de diferentes maneiras: para valores muito pequenos de p (valor de x_i pequeno) a um acréscimo alto na probabilidade e há um decrescimento até atingir patamares pequenos de variação de probabilidade em valores de p próximos de 1.

No gráfico a seguir, a curva azul representa a função $f(x) = \frac{1}{1+e^{x-5}}$ e a curva verde representa $f(x)$ depois da variação do coeficiente x_i em uma unidade, isto é, representa a função $g(x) = \frac{1}{1+e^{x-5}e^{-\beta_i}}$, aqui tomamos um valor de exemplo $\beta_i = 1.5$. A curva tracejada laranja representa a função $\frac{g(x)-f(x)}{f(x)}$ que é a comparação da variação entre as duas curvas, assim como deduzimos, observamos que nos valores pequenos de x a variação é grande e há uma diminuição desta variação ao chegar nos valores altos de x.

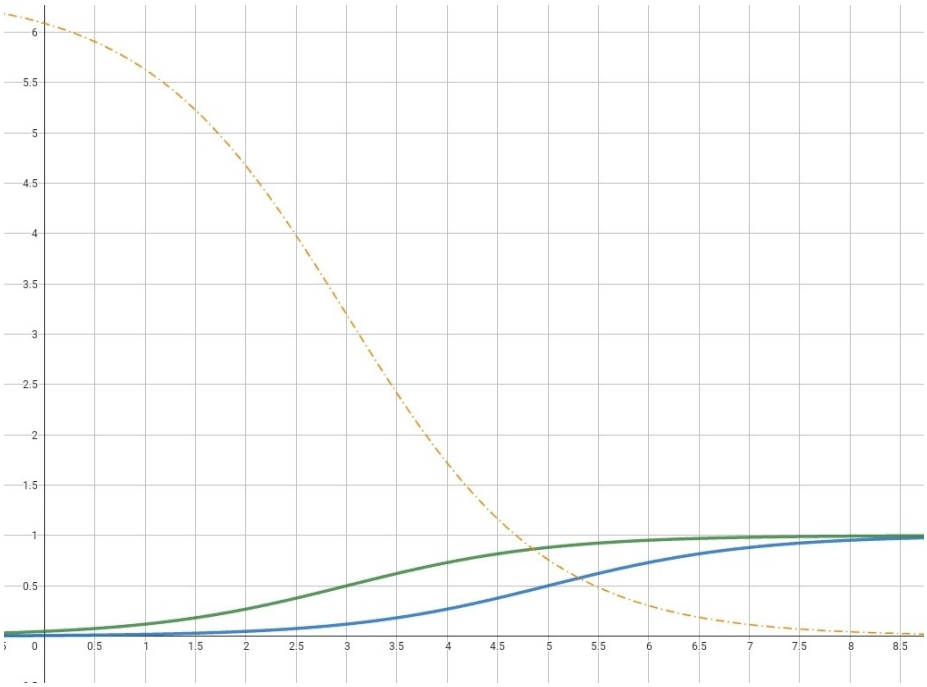


Figura 2.2: *Comparação gráfica entre $f(x)$ e $g(x)$. Fonte: produção com software Geogebra*

2.3.4 Critérios de Informação

Critérios de informação são métricas utilizadas para comparação de modelos de forma objetiva: se por um lado o aumento das variáveis que são incorporadas no modelo implica em uma maior capacidade de captar mais informações relevantes para predizer a variável alvo, por outro lado pode levar a um *overfitting*, que ocorre quando o modelo se ajusta tão bem aos dados de treino que é incapaz de se adequar aos dados de teste, resultando em um modelo ineficaz.

É justamente na busca de se equlibrar simplicidade e precisão que existem diversos critérios de informação na literatura, dentre os que mais se destacam são o AIC (Akaike Information Criterion) e o BIC (Bayesian Information Criterion) que são calculados por:

$$AIC = -2log(l(\beta)) + 2k$$

$$BIC = -2log(l(\beta)) + klog(n)$$

onde k é o número de parâmetros do modelo.

É facil notar a partir dessas equações que no BIC há uma penalização mais rigorosa para modelos mais complexos do que o AIC devido a termo $log(n)$. Tendo em vista que desejamos formular modelos mais interpretáveis possíveis, modelos com sufiente verossimilhança e menor quantidade de variáveis são preferíveis, logo o critério de informação que será utilizado para escolha da melhor regressão linear dos dados será o BIC, cujo referencial teórico pode ser encontrado em Andrew A. Neath, 2011.

2.3.5 Avaliação das significâncias dos parâmetros

No geral, a literatura cita dois testes para a avaliação da significância dos parâmetros do modelo de Regressão Logística: Teste de Wald e o Teste da Razão de Verossimilhança. Aqui vamos nos limitar ao primeiro, o leitor interessado pode se aprofundar em ambos em: UFRN, 2021.

O teste de Wald testa a hipótese nula que $H_0 : \beta_j = 0$ contra a alternativa de que $H_1 : \beta_j \neq 0$, utilizando a seguinte estatística:

$$W_j = \frac{\beta_j}{\hat{\sigma}(\hat{\beta}_j)}$$

Sob a hipótese inicial que $y_i \sim \text{Bernoulli}(p_i)$, onde $p_i = \frac{1}{1+e^{-g(X_i,\beta)}}$ juntamente com a hipótese nula, W_j possui distribuição assintótica qui-quadrado com 1 grau de liberdade. Rejeitamos H_0 caso a probabilidade da cauda associada a W_j seja menor em módulo do que o nível de significância α.

2.4 Árvore de Decisão

2.4.1 Introdução

Na abordagem da árvore de decisão temos as variáveis indepedentes $x_1, x_2, ..., x_k$ formando um subespaço $X \subset \mathbb{R}^k$ que será usado como base para predizer o valor da variável binária Y a partir da divisão desse subespaço. Então, supondo que iremos subvidir o subespaço em w regiões, para cada região R_j será mais provável Y assumir o valor 0 ou o valor 1.

Temos aqui a principio dois problemas: o primeiro é como será feita a divisão entre as regiões e o segundo é qual é o número de regiões ideais para os nossos dados. Em relação ao primeiro problema, iremos utilizar a técnica de *recursive binary splitting*, que consiste em começar escolhendo aleatoriamente um preditor x_j e um ponto de corte s_j que irão definir duas subregiões: $\mathbb{R}_1 = \{X|x_j < s_j\}$ e $\mathbb{R}_2 = \{X|x_j \geq s_j\}$.

Depois disso, para avaliar a qualidade do corte utilizamos o índice de Gini que mede o grau de pureza de cada corte j, onde K é o número de possíveis classes (em nosso caso são duas).

$$G = \sum_{k=1}^{K} \hat{p}_{jk}(1 - \hat{p}_{jk})$$

Isto é, valores extremos de G indicam predominância de uma classe, enquanto que o contrário indica um maior equlíbrio das classes ao longo da região. Dentro do nosso contexto de classificação, o primeiro cenário é preferível, pois dado que uma observação irá pertencer a uma região, se ela estiver fortemente associada a apenas uma classe, temos evidências para concluir com alta probabilidade que o Y relativo a esse caso pertence a classe predominante.

Então iremos proceder da seguinte maneira: na primeira iteração iremos escolher um preditor x_j e um ponto de corte s_j que minimizam G, após isso duas subregiões serão criadas. Para cada nova região, iremos buscar outro preditor x_i e um ponto de corte s_i que minimizam G e assim, sucessivamente...

2.4.2 *Overfitting*

Em relação ao segundo problema, é natural pensarmos em estender o número de regiões w ao infinito para obter maior precisão na classificação, entretanto se o fizermos irá ocorrer um super-ajuste do modelo aos dados de treino, absorvendo características próprias desse conjunto de dados. O grande problema é que o modelo acaba se adaptando tanto aos dados de treino que não consegue ter uma eficiência razoável em outros conjuntos de dados, é o que chamamos de *overfitting* do modelo.

Existem algumas alternativas que são adotadas para evitar o *overfitting*, como por exemplo: limitar a profundidade da árvore, definir uma redução minima na impureza para dividir uma região, dividir o conjunto de treino em vários subconjuntos para treinar o modelo, ou utilizar o *Random Forest*, que é a técnica que iremos utilizar neste projeto.

2.4.3 *Random Forest*

A idéia do processo de Random Forest é utilizar T árvores de decisão independentes entre si para melhor embasar a decisão preditiva. Aqui vamos tratar cada árvore como uma função f que recebe como argumentos o vetor de variáveis independentes X e retorna, no caso específico da classificação binária, o valor 0 ou 1:

$$f_T : \mathbb{R}^k \to \{0, 1\}$$

Cada árvore recebe como argumento uma amostra aleatória X distinta, evitando haver super-ajuste para uma amostra específica. Então os procedimentos são feitos semelhante ao que comentamos na primeira subseção e a decisão da floresta é definida pelo valor de saída de f mais recorrente dentre as T árvores:

$$\hat{Y} = \text{mode}\left(f_1(X), f_2(X), \ldots, f_T(X)\right)$$

A figura esquemática a seguir representa o processo descrito nesta subseção, onde as classes de bons e maus pagadores são representadas pelas cores verde e vermelho, respectivamente.

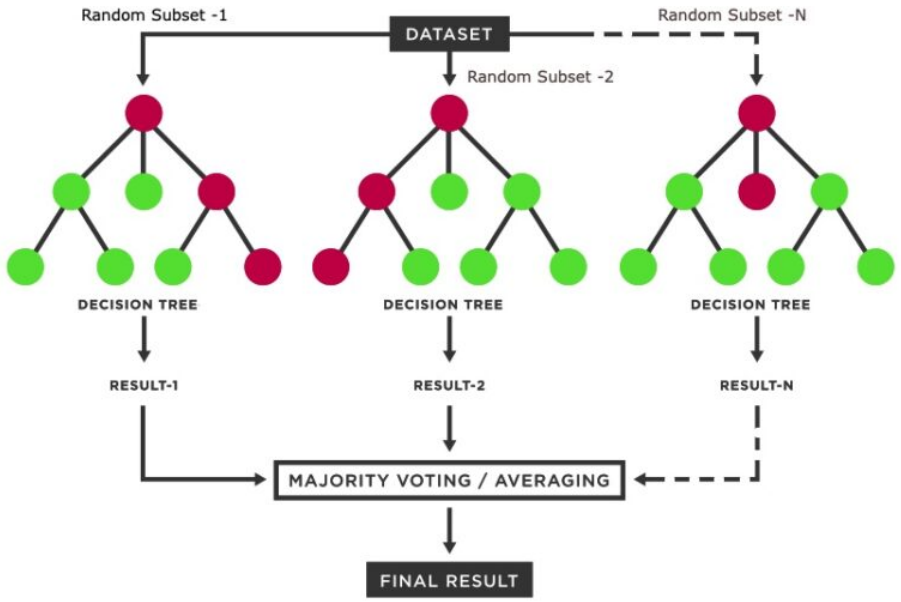


Figura 2.3: *Diagrama de Random Forest. Fonte: INFORYAN - Research & Data Consulting)*

Capítulo 3

Extração e Tratamento dos Dados

3.1 Fonte de Dados

A obtenção de dados de crédito de forma gratuita no Brasil sempre se mostrou um desafio, por um lado o funcionamento deste mercado faz com que a aquisição dessas informações seja permitida apenas através de bureaus de crédito, que em sua grande maioria possuem alto de custo de consulta. Por outro lado, leis de proteção de dados como a LGPD impoem regras rígidas sobre coleta, armazenamento e compartilhamento dessas informações.

Devido a estas restrições, os dados utilizados nesse estudo foram obtidos da plataforma digital Kaggle e possuem informações de renda anual, prazo do empréstimo, dentre outras, relativas a empréstimos pessoais sem garantia que foram realizados através de uma plataforma online nos Estados Unidos durante período de 2007 a 2018.

Os dados foram obtidos a partir da empresa de serviços financeiros Lending Club que até o ano de 2020 possuia uma plataforma de empréstimos onde era possível o usuário criar um anúncio do seu empréstimo a fim de atrair investidores interessados no financiamento deste. Na outra ponta, a plataforma permitia que os investidores visualizassem informações de risco de crédito da operação com o mutuário para auxiliar na decisão de investir ou não naquela aplicação, assim como pode ser visto em um *p*rint do *s*ite da plataforma a seguir.

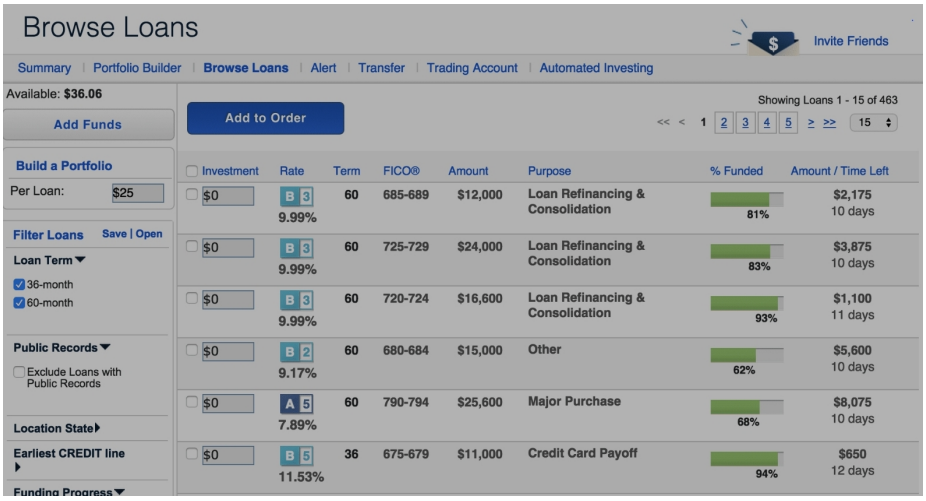


Figura 3.1: *Plataforma do Lending Club. Fonte: Fintch Nexus*

No período de coleta da base de dados (entre 2007 e 2018), os propósitos de empréstimos nos Estados Unidos refletiram diferentes demandas econômicas e sociais do período, moldadas pela crise financeira de 2008 e a recuperação subsequente. De acordo com dados do *C*FPB, *L*ending Tree e *TransUnion* extraídos com auxílio do *ChatGPT*, os empréstimos hipotecários para compra de imóveis representaram até 65% dos novos financiamentos no final do período, impulsionados por taxas de juros mais baixas e regulamentações aprimoradas. Empréstimos pessoais também tiveram destaque, sendo utilizados principalmente para consolidação de dívidas (39,2%) e refinanciamento de cartões de crédito (21,8%), enquanto uma parcela menor foi destinada a melhorias residenciais (7,7%) e gastos médicos (3,0%).

A tabela a seguir contém as proporções de propósitos para os empréstimos encontrados na base de dados e é possível verificar que a maioria dos clientes contrataram empréstimo para consolidação de dívidas, o que consideramos um perfil mais arriscado, pois são clientes que já estão em processo de dívidas com outras instituições.

Purpose	Descrição	Proporção
`debt consolidation`	consolidação de dívidas	56.53%
`credit card`	consolidação de dívida de cartão de crédito	22.87%
`home improvement`	reformas na casa	6.66%
`major purchase`	compra de alto valor	2.23%
`medical`	gastos médicos	1.22%
`small business`	capital para pequenos negócios	1.09%
`car`	compra ou manutenção de automóvel	1.06%
`others`	outros propósitos	8.35%

Tabela 3.1: *Proporções de propósitos dos empréstimos na base de dados original*

3.2 Refinamento da Base

O principal desafio encontrado ao utilizar esta base foi o de processamento de dados, já que estamos utilizando a versão gratuita do software Google Colab que não possui

recursos suficientes para processar as dimensões de mais de 2 milhões de linhas e 145 variáveis que compõe o nosso conjunto de dados.

Devido a essa restrição, escolhemos por particionar a base de acordo com a variável *loan status* que assume os valores da tabela abaixo. Vemos que para ela existem três classficações muito representativas (empréstimo completamente pago, empréstimo em andamento e empréstimo não pago) e outras classificações minoritárias relativas a faixas de rolagens nos atrasos. Por uma questão de simplicidade e comparação de métodos, escolhemos nos restringir apenas aos casos em que o tomador pagou completamente ou não o empréstimo, representados por *Fully Paid* e *Charged Off*.

Loan status	Descrição	Proporção
`Fully Paid`	empréstimo totalmente pago	46.09%
`Current`	empréstimo ativo	40.68%
`Charged Off`	empréstimo não pago sem expectativa de pagamento	11.57%
`Others`	todas outras classificações	1.66%

Tabela 3.2: *Proporções de empréstimos na base de dados original*

Assim como feito em Gonçalves *et al.*, 2013, separamos os dados em um conjunto de treino para os modelos, composto por 175000 bons pagadores e 175000 maus pagadores selecionados aleatoriamente dos dados originais. Analogamente, teremos um conjunto de dados para testes dos modelos que será composto por 75000 maus pagadores e 75000 bons pagadores.

3.3 Tratamento das Variáveis

O primeiro passo para construirmos bons modelos preditivos compreende ter uma base de dados de treino consistente. Por isso uma investigação inicial no conjunto de dados foi realizada para identificar problemas que possam atrapalhar o desempenho dos modelos e foram aplicados os seguintes metódos de contorno baseados nos capítulos 7 e 8 de McKinney, 2018:

- **Variáveis do tipo texto**: As bibliotecas do Python para modelos preditivos (*Skitlearn* e *Scipy*) não se comportam bem com variáveis do tipo texto, por isso transformamos as 36 variáveis desse tipo encontradas na base para o tipo numérico, adotando a abordagem de atribuir um número para cada categoria.

- **Variáveis com observações nulas:** Algumas variáveis apresentaram observações nulas, o procedimento adotado foi utilizar a função *fillna()* da biblioteca *Pandas* para substituir essas observações por zero. Pensando na óptica dos modelos que vamos aplicar aos dados, não há problema em proceder dessa maneira, pois estes descartam naturalmente variáveis que possuem valores em comum para observações que possuem valores distintos de variável target (*loan status*).

- **Variáveis de identificação:** Resolvemos por excluir do conjunto de dados variáveis relativas a identificação dos mutuários (*id, member id, url*), já que no contexto de

previsão de bons ou maus pagadores, estas são irrelevantes para o modelo na medida em que não trazem nenhuma informação adicional de propensão à inadimplência.

- **Variáveis não padronizadas:** Algumas variáveis como a descrição do objetivo do empréstimo (*desc*) apresentaram um valor muito grande de classes com pouco representatividade, portanto foi decidido utilizar uma classificação binária em que a variável recebe o valor 1 se o campo estiver preecnhido e 0 caso contrário.

3.4 Preparação dos Dados

Devido à existencia de uma alta quantidade de variáveis (145) na base de dados, optamos também por aplicar métodos de redução de variáveis que carregavam consigo informações não agregadoras aos modelos ou informações já expressas em outras variáveis.

A primeira abordagem foi identificar grupos de variáveis que fossem muito correlacionadas entre si, já que uma alta correlação entre duas variáveis é indicativo de que essas duas variáveis irão agregar as mesmas informações para o modelo, ou ainda distorcer resultados deste, como Joseph F. Hair *et al.*, 2005 comenta no caso da Regressão Logística. Primeiramente calculamos a matriz C $145x145$ de correlação entre as 145 variáveis, onde cada entrada é dada por $c_{ij} = corr(X_i, X_j)$ onde $X_1, X_2, ..., X_{145}$ são as variáveis contidas na base. Para cada par de variáveis consideramos a "distância"entre essas duas variáveis como:

$$d_{ij} = 1 - |corr(X_i, X_j)|$$

Como nosso objetivo é identificar as variáveis que possuem menor "distância"entre si, foi utilizado o algoritmo de agrupamento hierarquico Linkage cujo referencial teórico pode ser consultado em Hastie e Friedman, 2009. Em linhas gerais, o algoritmo se inicia atribuindo a cada variável um cluster e a cada iteração o procedimento de comparar a distância entre os clusters é feito e caso haja dois clusters com distância menor do que o *threshold* previamente determinado, esses são transformados em um só cluster e o processo se desenrola até a distância minima entre quaisquer dois clusters ser maior que o *threshold*.

O critério utilizado como *threshold* foi de 0.9 e a aplicação do Linkage na base de dados retornou 32 clusters com mais do que duas variáveis, de maneira que variáveis que estão no mesmo cluster irão ter contribuições semelhantes de informação em nosso modelo. Tendo por objetivo a redução do número de variáveis, foi adotado o procedimento de escolha de uma variável representativa de cada cluster com mais do que uma variável, o critério de escolha foi baseado em optar preferencialmente por variáveis mais citadas na literatura e de fácil entendimento do público geral.

A segunda abordagem utilizada a fim de compactar os dados foi excluir da base variáveis que são observadas apenas após o fechamento ou evolução do contrato, dentre elas estão: pagamento total até o momento, último pagamento, valor total recuperado em caso de perda, dentre outras. Após essa etapa foram eliminadas 14 variáveis da base de dados.

A terceira abordagem foi a utilização do algoritmo de *Random Forest* para estimar a importância de cada variável do conjunto de dados na predição da variável *loan status*, assim como descreve Biau e Scornet, 2016 na seção 5. Em suma, o procedimento realizado

é o teste de redução da impureza média dos nós de cada árvore devido a exclusão de uma variável, lembre-se de que quanto mais extremo o grau de pureza, maior a assertividade do modelo, então se ao retirar uma variável o grau de impureza médio não diminui/aumenta consideravelmente, há o indicativo de que aquela variável não é tão importante para o modelo. As variáveis com os 10 maiores graus de importância estão listadas abaixo e todas as demais foram excluídas a fim de possuirmos uma base mais refinada para proseguir na aplicação das técnicas.

Variável	Descrição	Importância
`int rate`	A taxa de juros do empréstimo	8.53%
`dti`	A relação dívida/renda do mutuário	4.34%
`bc open to buy`	Crédito disponível em linhas rotativas	3.59%
`revol bal`	O saldo total de crédito rotativo do mutuário	3.56%
`avg cur bal`	Saldo médio atual de todas as contas	3.55%
`loan amnt`	O valor do empréstimo solicitado	3.55%
`annual inc`	A renda anual do mutuário	3.55%
`tot cur bal`	Saldo total atual de todas as contas	3.39%
`bc util`	Utilização da linha de crédito atual	3.36%
`total bc limit`	Total de limites de crédito do Banco	3.35%

Tabela 3.3: *Dez variáveis com melhores reduções na impureza dos nós no Random Forest*

3.5 Remoção de outliers e normalização dos dados

A princípio o conjunto de dados pode conter em determinadas variáveis observações com valores muito baixos ou muito elevados para o comportamento médio da variável, esses valores são chamados de outliers e geralmente são fruto de erros de imputação de dados ou da ocorrência de casos exóticos. Muitos algoritmos preditivos são prejudicados pela presença de outliers, pois esses valores extremos acabam distorcendo a distribuição dos dados e forçando o modelo a se adaptar a esses poucos valores extremos ao invés de se adaptar ao padrão geral dos dados. Seguindo a referência de Hastie e Friedman, 2009 , estaremos utilizamos como base o intervalo interquartil para eliminação de outliers encontrados na base de dados.

O processo de normalização dos dados também é importante pois algumas técnicas preditivas que atribuem pesos às variaveis (como a Regressão Logística), podem ser afetadas negativamente se as variáveis não estiverem em uma mesma escala. Para verificar a suposição de normalidade dos dados foi utilizado o teste de Shapiro-Wilk que é um teste não paramétrico que tem por objetivo testar a hipótese nula de que a população dos dados que geraram uma amostra tem uma distribuição normal, mais detalhes podem ser encontrados em Malato, 2023.

Aplicação do teste em todas as variáveis nos revelou que nenhuma seguia uma distribuição normal, então a fim de não perdermos a distribuição original dos dados, utilizamos como sugerido em Codecademy, 2024 a normalização *max-min* que mantém a distribuição das variáveis com valores entre 0 e 1.

Capítulo 4

Desenvolvimento do Modelo de Risco

4.1 Motivação

Os algoritmos que vimos para decisão da concessão de crédito como a Regressão Logística e Árvore de Decisão no geral possuem bom desempenho, porém uma das desvantagens que apresentam é que não englobam o sentido do risco de crédito, isto é: para esses modelos, classificar um mutuário com probabilidade alta de inadimplência como possível cliente tem o mesmo peso do que classificar um mutuário com baixa probabilidade de inadimplência como não possível cliente, o que é conceitualmente equivocado no ponto de vista de crédito, pois um cliente com alta probabilidade de inadimplência tende a trazer mais prejuízo financeiro em relação a um mutuário com bom comportamento financeiro que não foi captado pelo algoritmo. Esse ponto de vista também é compartilhado por Casa Nova, 2013 e é um dos principais motivadores para o desenvolvimento do modelo.

4.2 Escolha das Variáveis

Observando a ampla citação na literatura sobre quão determinante é a capacidade de renda na previsão de inadimplência futura e também a utilização recorrente por vários bureaus de crédito (como exemplo o score FICO, cuja descrição mais detalhada se encontra em FSF, 2024), tomamos aqui a liberdade de criar uma variável que representa a capacidade de pagamento do empréstimo, que é definida por:

$$I = \frac{V\dot{(}1 + i)^{60}}{60R}$$

Onde V é o valor do empréstimo em dólares, i a taxa de juros e R a renda mensal do mutuário em dólares. Tendo em vista que a maioria dos empréstimos da base são de prazo 60 meses, esta fórmula em termos financeiros estima a proporção da renda mensal do individuo que será destinada ao pagamento médio mensal das parcelas do empréstimo.

O primeiro critério escolhido para estimar a importância das variáveis para compor o modelo foi baseado em um estudo com testes de hipóteses realizado com o objetivo de

identificar quais dentre as 10 variáveis obtidas do produto final da base descriminavam melhor a população que apresentou histórico de pagamento desejável em relação a população dos demais pagadores. Como já sabemos da seção 3.5 que os dados não seguem uma distribuição normal, optamos por utilizar aqui o teste não paramétrico de Mann-Whitney U para verificar se sobre aquela variável especifica, as populações de bons e maus pagadores tinham médias diferentes, para o leitor interessado, um aprofundamento sobre o teste pode ser visto em McClenaghan, 2024.

Após aplicação dos testes, em todas as variáveis houve significância estatística (considerando $\alpha = 5\%$) para afirmar que as variáveis distinguiam bem (em quesito de média) as duas populações, o que é um bom indicativo, porém não há critérios de escolhas de variáveis mais importantes por este teste.

Como segundo critério de decisão de variáveis para o modelo, utilizamos conjuntamente a importância obtida na etapa de *Random Forest* com a quantidade de recorrências na literatura que foi revisada durante este trabalho. As 3 variáveis escolhidas estão listadas abaixo, porém é importante salientar que mais variaveis podem ser inclusas no modelo para melhorar seu desempenho.

Variável	Descrição
`Index`	Indice de Comprometimento de renda
`dti`	Relação dívida/renda do mutuário
`avg cur bal`	Saldo médio atual de todas as contas do mutuário

Tabela 4.1: *Variáveis utilizadas na construção do modelo de risco*

4.3 Pontos de Cortes das Variáveis

Atualmente no mercado financeiro muitos algoritmos de decisão são utilizados sem o devido cuidado de interpretação do seu processo de decisão e de seus *outputs* finais. Observando essa problemática, optamos por utilizar um modelo baseado em árvore de decisão no qual precisamos definir para cada variável um ponto de corte que determinará se o mutuário seguirá no fluxo de contratação do empréstimo ou será descartado. Portanto, o mutuário deverá passar pelos 3 pontos de decisão envolvendo as variáveis anteriormente citadas para poder ser aprovado na concessão do empréstimo.

O primeiro passo para definir os pontos de corte das variáveis passa por entender qual será o objetivo do modelo. A princípio, pensou-se aqui em um modelo que pudesse servir como parâmetro para escolher em quais clientes vale assumir o risco de conceder um empréstimo, observando a probabilidade desses clientes se tornarem inadimplentes. O conceito de risco pressupõe a perda financeira e em termos comparativos, é muito mais custoso deixar de lucrar com um cliente que honrará seus compromissos do que ter que arcar com o prejuízo de perdas de clientes inadimplentes, portanto vamos definir aqui dois tipos de erros que nosso modelo poderá cometer:

- Erro Tipo I : Consiste em aceitar a concessão de empréstimo a um mutuário que futuramente deixará de cumprir as suas obrigações com a instituição;

- Erro Tipo II: Consiste em não aceitar conceder o empréstimo a um mutuário que futuramente não deixaria de cumprir as suas obrigações com a instituição.

Em nosso contexto, é desejável que o erro tipo I represente um percentual pequeno do total de erros cometidos pelo nosso modelo. Para uma primeira simulação, estaremos considerando que o erro tipo I não ultrapasse a proporção de um terço do total de erros cometidos. Então a primeira etapa foi considerar a decisão de concessão ou não do empréstimo apenas pela variável *index*. O ponto de corte obtido foi de aproximadamente 2.5 através do seguinte algoritmo:

- Etapa I : Escolher um ponto de corte inicial para a variável *index* a partir de um valor entre entre as medianas dos maus e bons pagadores para essa variável. Em nosso caso, utilizamos $index = 15.0$ como estimativa inicial;
- Etapa II: Escolher da base de treino uma amostra aleatória de 1000 bons pagadores e 1000 maus pagadores;
- Etapa III: Para o ponto de corte da iteração verificar se o erro tipo 1 representa menos que um terço do total de erros. Em caso positivo, este será o ponto de corte e em caso negativo reduz-se o ponto de corte em -0.5 até encontrar o porte de corte que satisfaça essa condição;
- Etapa IV: Repete-se as etapas I, II, III 1000 vezes e tira-se a média dos pontos de cortes obtidos.

Considerando agora apenas a subpopulação com $index < 2.50$, isto é, que seria aprovada na primeira versão de nosso modelo, simulamos o mesmo algoritmo para a variável dti, obtendo o valor de corte de aproximadamente 20.98. Agora considerando a subpopulação tal que $index < 2.5$ e $dti < 20.98$, após aplicação novamente do algoritmo obtemos um valor de corte de 12487.50 para a variável $avgcurbal$. Após a definição dos pontos de corte, o algoritmo final ficou como esquematizado na figura a seguir.

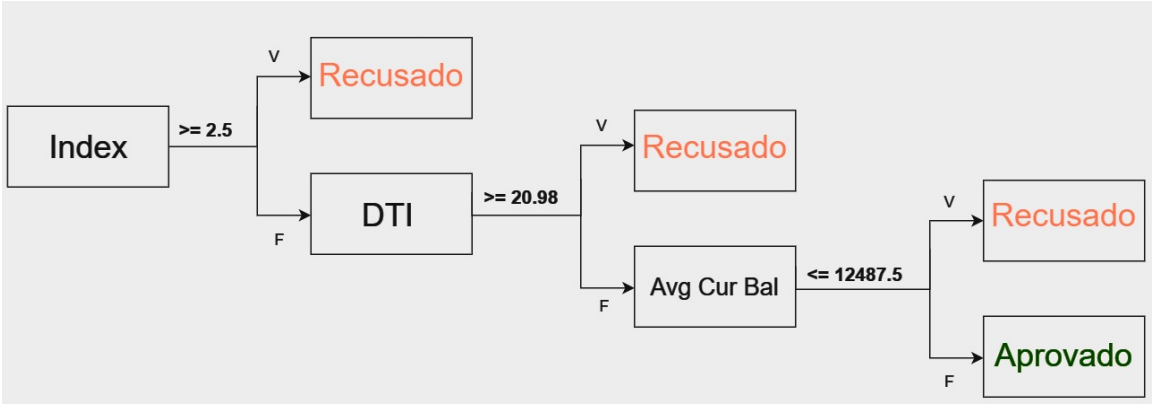


Figura 4.1: *Esquematização gráfica do modelo de risco (produção própria)*

4.4 Parâmetro de Flexibilidade

No modelo anterior, a razão máxima escolhida entre a quantidade de erros tipo I e os erros totais foi de 1/3. Entretanto, este parâmetro de flexibilidade poderia ser alterado

para um valor maior ou menor, a depender do critério de risco. Valores muito pequenos deste parâmetro indicam que o modelo é pouco tolerante ao erro tipo I, isto é possui uma decisão mais conservadora ao avaliar os clientes, enquanto que modelos com parâmetros de flexibilidade altos são mais abertos no sentido de apetite de risco.

A definição deste parâmetro pode ser muito complexa, porém para simplificar, sugerimos a comparação entre a quantia que seria ganha a partir dos juros dos empréstimos de clientes bons (considerando aqui alguma medida de ticket e taxa de juros média dentro deste grupo) e a quantia perdida com os clientes inadimplentes (também considerando um perfil médio). Como a priori não temos uma definição de grupo de clientes bons ou ruins pela falta de clientes na carteira, podemos considerar um parâmetro mais baixo no ínicio das operações e ir calibrando-o conforme tivemos maior população de clientes que tiveram perdas e mutuários que foram recusados pelo algoritmo, calculando qual taxa e ticket teriam caso fechassem com a instituição.

A fim de testar empiricamente a mudança do parâmetro de flexibilidade, realizamos 3 simulações variando seu valor e registramos as mudanças nos pontos de cortes e na taxa de aprovação da base de clientes, os resultados são mostrados a seguir.

Flexibilidade	Corte *index*	Corte *DTI*	Corte *avg cur bal*	taxa de aprovação
0.25	1.50	23.48	15984.00	6.46%
0.33	2.50	20.98	12487.50	8.45%
0.50	6.99	17.98	8491.50	11.19%

Tabela 4.2: *Simulações variando parâmetro de flexibilidade*

A partir da análise da tabela, podemos ver que a variação do parâmetro de flexibilidade implica também em uma variação na taxa de aprovação, de maneira que em simulações com parâmetros altos, são esperadas taxas de aprovações altas e vice-versa. Além disso, outro ponto que podemos notar foi a variação também dos pontos de corte de acordo com o valor do parâmetro utilizado, acreditamos que essa característica do modelo é a principal diferenciadora em relação aos modelos tradicionais frente a um mercado onde empresas cada vez mais querem simular diferentes cenários de apetite de risco de crédito.

Capítulo 5

Avaliação e Aplicações

5.1 Objetivos do Capítulo

Este capítulo tem por objetivo realizar uma avaliação comparativa entre o modelo produzido neste trabalho e os modelos tradicionais utilizados no mercado de crédito (Regressão Logística e *Random Forest*), além de potenciais aplicações para casos que não são atendidos pelos modelos tradicionais.

Em seu artigo, Gonçalves *et al.*, 2013 utiliza o produto entre a taxa de acerto entre os bons pagadores e o maus pagadores como medida de avaliação de seu modelo logístico, justificando-o pelo fato de não haver informações *a priori* sobre o que seria mais atrativo para a instituição financeira (identificação de bons ou maus clientes). Em nosso contexto, temos claro que é o mais atrativo seria a identificação de maus clientes, entretanto não devemos simplesmente ignorar a identificação de bons pagadores. A fim de refletir melhor sobre essa questão, na seção a seguir iremos nos aprofundar nas métricas de performance de modelos.

5.2 Matriz de Confusão e suas métricas

De acordo com Müller e Guido, 2016 um dos meios mais compreensíveis para a interpretação dos resultados de uma classificação binária é utilizando uma matriz de confusão, que consiste em uma matriz M 2x2 onde as linhas correspondem às classes verdadeiras e as colunas representam as classes preditas. Em nosso contexto, as linhas terão rótulos indicando se o empréstimo foi quitado ou não, enquanto que as colunas terão rótulos indicando se o mutuário foi rejeitado ou não pelo modelo.

$$M = \begin{bmatrix} vn & fp \\ fn & vp \end{bmatrix}$$

Então, a partir da matriz podemos fazer a interpretação de que dentro dos mutuários que não pagaram totalmente suas dívidas, vn representa a quantidade de mutuários que foram rejeitados pelo modelo, enquanto que fp representa a quantidade de mutuários que

foram aceitos pelo modelo. De maneira análoga, dentro dos mutuários que conseguiram pagar totalmente seus empréstimos, fn representa a proporção que foi rejeitado pelo modelo, enquanto vp representa a proporção que foi aceita pelo modelo.

Embora a matriz seja uma boa forma de traduzir a classificação, a análise de matrizes para vários modelos simultâneamente acaba sendo cansativa e por isso são definidas métricas de resumo a partir de operações com seus elementos. Apresentamos aqui as quatro principais para nosso contexto:

- **Acurácia:** Representa o proporção de previsões corretas do modelo. É calculada por $(vn + vp)/T$, onde $T = vn + fp + fn + vp$ é o total de mutuários classificados pelo modelo;
- **Precisão:** Representa a proporção de mutuários aceitos pelo modelo que de fato conseguiram pagar os seus empréstimos. É calculada por $vp/(vp + fp)$;
- **Sensibilidade:** Representa a proporção de mutuários que pagaram completamente seus empréstimos e foram aceitos pelo modelo. É calculado por $vp/(fn + vp)$;
- **F1-Score:** É a média harmônica da precisão e da sensibilidade. Essa métrica é útil quando há um equilíbrio necessário entre minimizar falsos positivos e falsos negativos, mas pode ser ajustada se um tipo de erro for mais importante.

$$\text{F1-score}_i = 2 \times \frac{\text{Precisão}_i \times \text{Revocação}_i}{\text{Precisão}_i + \text{Revocação}_i}$$

Devido ao claro desbalanceamento das classes considerando o prejuízo financeiro à empresa, acreditamos que a métrica mais parcimoniosa a fim de obter uma comparação justa entre os modelos é o F1-Score ponderado, definido por:

$$\text{F1-score ponderado} = \sum_{i=0}^{n-1} w_i \times \text{F1-score}_i$$

Em nosso contexto, teremos uma classificação binária com $n = 2$ e, além disso, estamos considerando um peso de $1 - \alpha$ para a identificação correta de possíveis maus pagadores e de α para a identificação correta de possíveis bons pagadores, portanto temos os seguintes valores de w:

$$\begin{cases} w_0 = 1 - \alpha \\ w_1 = \alpha \end{cases}$$

5.3 Implementação dos modelos tradicionais

A fim de comparação de resultados com nosso modelo, implementamos também no ambiente Google Colab os modelos de Regressão Logística e *Random Forest*. A implementação do algortimo de Floresta de Decisão consistiu na previsão da variável *loan status*

pelas dez variáveis que chegamos ao final do capítulo 3, já que não há uma influência ativa entre variáveis neste modelo devido a garantia de baixa correlação entre elas.

No caso específico da Regressão Logística, o modelo é mais sensível à presença ou não de uma variável, já que os coeficientes se adequam à todas aos valores observados de todas variáveis. Então devemos tratar os $2^{10} - 1$ modos possíveis de inserção das variáveis como modelos distintos. Para escolher o modelo representativo optamos por utilizar o critério de informação BIC apresentado no capítulo 2 por dois motivos: primeiro por ser um método que consegue estimar uma medida quantitativa para comparar a verossimilhança dos modelos e o segundo por penalizar com maior rigor modelos com muitas variáveis, já que estamos em um cenário em que queremos um modelo que seja o mais interpretável possível por parte do usuário e a presença excessiva de variáveis pode nos atrapalhar nesse objetivo.

O procedimento adotado para escolha do menor BIC foi o seguinte: inicialmente todas as 10 variáveis candidatas foram alocadas em uma lista e um modelo de regressão logística foi declarado. Um loop foi iniciado sobre a lista e a cada iteração, era verificado se a inserção de dada variável diminuiria ou não o BIC do modelo, em caso positivo a variável era adicionada ao modelo e em caso negativo descartada, o processo foi realizado até a lista conter apenas variáveis que aumentariam o BIC do modelo. As variáveis selecionadas estão descritas na tabela abaixo:

Variável	Descrição
`int rate`	A taxa de juros do empréstimo.
`dti`	Relação dívida/renda do mutuário
`avg cur bal`	Saldo médio atual de todas as contas do mutuário
`loan amnt`	O valor do empréstimo solicitado
`total bc limit`	Total de limites de crédito do Banco
`annual inc`	A renda anual do mutuário

Tabela 5.1: *Variáveis selecionadas para o modelo de menor BIC*

Adicionalmente também foram coletados os valores dos coeficientes de cada variável da regressão, bem como as estatíscas e os p-valores do teste de Wald (apresentado no capítulo 2), pelo qual chegamos a conclusão de que todas as variáveis apresentadas na tabela anterior e o intercepto são estatisticaente significantes para o modelo.

```
Teste de Wald (Significância dos Coeficientes):
                          Logit Regression Results
==============================================================================
Dep. Variable:            loan_status   No. Observations:               317580
Model:                          Logit   Df Residuals:                   317573
Method:                           MLE   Df Model:                            6
Date:                Fri, 22 Nov 2024   Pseudo R-squ.:                0.08071
Time:                        15:08:34   Log-Likelihood:            -2.0227e+05
converged:                       True   LL-Null:                   -2.2003e+05
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.2547      0.015     83.622      0.000       1.225       1.284
int_rate      -3.0127      0.023   -130.065      0.000      -3.058      -2.967
dti           -1.2041      0.029    -41.341      0.000      -1.261      -1.147
avg_cur_bal    0.8593      0.022     39.712      0.000       0.817       0.902
loan_amnt     -0.9118      0.022    -40.881      0.000      -0.955      -0.868
total_bc_limit 0.5460      0.023     23.336      0.000       0.500       0.592
annual_inc     0.4318      0.030     14.404      0.000       0.373       0.491
==============================================================================
```

Figura 5.1: *Resultados do teste de Wald para regressão logística*

5.4 Resultados e discussão

Para cada modelo implementado neste projeto (Modelo de Risco variando o parâmetro de flexibilidade, Regressão Logística e Floresta Aleatória) calculamos as métricas de acurácia, precisão, sensibilidade (*recall*), F1-Score ponderado pelo parâmetro de flexibilidade (assim como mostrado no final da seção 5.2), além da taxa de aprovação da base de teste. Os resultados das simulações estão apresentados nas tabelas a seguir.

Simulação 1 com $\alpha = 0.25$					
Modelo	**Acurácia**	**Precisão**	**Sensibilidade**	**F1-Score**	**Aprovação**
Regressão Logística	63, 76%	62, 58%	62, 32%	64, 35%	48, 15%
Floresta Aleatória	67, 34%	73, 18%	51, 24%	69, 28%	33, 86%
Modelo de Risco	55, 14%	77, 09%	10, 29%	66, 56%	6, 45%

Tabela 5.2: *Tabela com resultados da simulação feita com* $\alpha = 0.25$

Simulação 2 com $\alpha = 0.33$					
Modelo	**Acurácia**	**Precisão**	**Sensibilidade**	**F1-Score**	**Aprovação**
Regressão Logística	63, 76%	62, 57%	62, 32%	64, 15%	48, 15%
Floresta Aleatória	67, 34%	73, 18%	51, 24%	68, 32%	33, 86%
Modelo de Risco	55, 84%	74, 82%	13, 07%	66, 82%	8, 45%

Tabela 5.3: *Tabela com resultados da simulação feita com* $\alpha = 0.33$

Como comentamos anteriormente, a medida mais parcimoniosa para comparação dos modelos é o F1-Score, observamos que a floresta aleatória no geral possui maior F1-score ponderado dentre todos os modelos, seguida do Modelo de Risco e da Regressão Logistica, portanto o *R*andom Forest continua sendo o modelo que lida melhor com o *trade-off* entre rejeitar um possível bom cliente e aceitar um possível mau cliente.

No quesito de acurácia, o modelo de Floresta Aleatória segue sendo o melhor, seguido pela Regressão Logística e por fim pelo Modelo de Risco. Como vimos no ínicio do capítulo,

Simulação 3 com $\alpha = 0.50$					
Modelo	**Acurácia**	**Precisão**	**Sensibilidade**	**F1-Score**	**Aprovação**
Regressão Logística	63, 76%	62, 57%	62, 32%	63, 72%	48, 15%
Floresta Aleatória	67, 34%	73, 18%	51, 24%	66, 28%	33, 86%
Modelo de Risco	56, 36%	70, 04%	17, 02%	66, 73%	11, 75%

Tabela 5.4: *Tabela com resultados da simulação feita com* $\alpha = 0.50$

essa métrica atribui o mesmo peso de classificar corretamente um possível bom cliente como apto a entrar na instituição e um possível mau cliente a ser rejeitado pela instituição, então como a construção do modelo não levou em conta a maximização do número de classificações corretas, mas sim uma compensação entre classificações corretas e perda financeira envolvida devido à essa decisão, é natural que tenhamos uma acurácia menor comparado aos outros modelos.

Quando olhamos para a métrica de precisão, notamos que os modelos com menor parâmetro de flexibilidadde possuem no geral a maior precisão dentre os três modelos comparados, isso reflete o conservadorismo do Modelo de Risco, já que este é muito criterioso na escolha de mutuários que irão ser aceitos. As métricas que mais chamaram atenção no quesito de distanciamento em relação às métricas dos modelos tradicionais foram a sensibilidade e a taxa de aprovação. Na construção do Modelo de Risco, optamos por tomar regras mais duras em cada ramo da àrvore, isto é, o mutuário precisaria satisfazer todas as regras simultanealmente para ser aprovado pelo modelo, devido a esse conservadorismo, é natural que o modelo não consiga capturar possíveis bons clientes que não satisfaçam pelo menos uma das regras, o que também acaba refletindo em uma taxa de aprovação abaixo da média dos modelos tradicionais.

Vemos então que é de grande validade o teste do Modelo de Risco em instituições que trabalham com concessão de crédito à clientes de alto risco. Embora o modelo não tenha expressiva taxa de aprovação, é necessário sempre ponderar os ganhos a curto prazo advindos da aprovação de clientes com possíveis perdas a longo prazo associada aos clientes que não conseguem honrar suas parcelas ao longo da vida do empréstimo.

Apêndice A

Códigos

Os programas utilizados nesse projeto foram desenvolvidos no ambiente do Google Colaboratory e podem ser encontrados em minha página pessoal do github, através do link.

O notebook $0RefinamentoBaseTCC$ tem por objetivo transformar a base bruta em bases menores de treino e teste para os modelos. Já o notebook $1TratamentoDadosTCC$ realiza o tratamento de todas as variáveis do tipo texto, transformando-as em numéricas ou excluindo-as caso necessário. Em seguida, o notebook $2AnaliseExploratoria$ é responsável por realizar os demais tratamentos nos dados: exclusão das variáveis altamente correlacionadas, redução da base, normalização e remoção de *outliers*.

Dentro do notebook $3ModeloRisco$ houve toda a construção do modelo proposto e o cálculo de suas respectivas métricas. Por fim, no notebook $4ModelosTradicionais$ os modelos de Regressão Logística e Floresta Aleatória são aplicados e suas métricas também calculadas.

A base de dados utilizada pode ser obtida através do link, lembre-se de baixar a base em arquivo .csv e salvá-la em seu drive. Após isso, é preciso atualizar os caminhos de todos os notebooks para a pasta que você criou no drive.

Referências

[Andrew A. Neath 2011] Joseph E. Cavanaugh Andrew A. Neath. "The bayesian information criterion: background, derivation, and applications". *Wires Computational Statistics* 4.2 (dez. de 2011), pp. 199–203. doi: https://doi.org/10.1002/wics.199 (citado na pg. 13).

[Biau e Scornet 2016] G. Biau e E. Scornet. "A random forest guided tour." *TEST* 25 (2016), pp. 197–227 (citado na pg. 20).

[Carvalho *et al.* 2007] Fernando J. Cardim de Carvalho, Francisco Eduardo Pires de Souza, João Sicsú e Luiz Fernando Rodrigues de Paula. *Economia Monetária e Financeira: Teoria e Política*. 1ª ed. Elsevier-Campus, 2007 (citado na pg. 4).

[Casa Nova 2013] Silvia Pereira de Castro Casa Nova. "Quanto pior, melhor: estudo da utilização da análise por envoltória de dados em modelos de análise de inadimplência/insolvência de empresas". *Revista Contemporânea de Contabilidade* 10 (2013), pp. 71–96 (citado na pg. 23).

[Codecademy 2024] Codecademy. *Normalization*. 2024. url: https://www.codecademy.com/article/normalization (acesso em 25/11/2024) (citado na pg. 21).

[D.R. e E.J. 1989] Cox D.R. e Snell E.J. *Analysis of Binary Data*. Chapman e Hall/CRC, 1989 (citado na pg. 10).

[FSF 2024] FICO. *What's in my FICO® Scores?* 2024. url: https://www.myfico.com/credit-education/whats-in-your-credit-score (acesso em 14/11/2024) (citado na pg. 23).

[Figueiredo 2001] R. P. Figueiredo. "Gestão de riscos operacionais em instituições financeiras – uma abordagem qualitativa". *Universidade da Amazônia, Belém – Pará* 1 (2001), pp. 9–10 (citado na pg. 7).

[Gastaldi 2005] J. Petrelli Gastaldi. *Elementos de economia política*. 1ª ed. Saraiva, 2005 (citado na pg. 3).

[Geeks For Geeks 2024] Geeks For Geeks. *Newton Raphson Method*. 2024. url: https://www.geeksforgeeks.org/newton-raphson-method/ (acesso em 14/11/2024) (citado na pg. 11).

[Gonçalves *et al.* 2013] Eric Bacconi Gonçalves, Maria Aparecida Gouvêa e Daielly Melina Nassif Mantovoni. "Análise de risco de crédito com o uso da regressão logística". *Revista Contemporânea de Contabilidade (RCC)* (2013) (citado nas pgs. 8, 19, 27).

[Hastie e Friedman 2009] R. Hastie T.and Tibshirani e J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009 (citado nas pgs. 20, 21).

[Joseph F. Hair *et al.* 2005] Jr. Joseph F. Hair, R.E. Anderson, R.L.Tatham e W.C. Black. *Analysis of Binary Data*. Bookman, 2005 (citado na pg. 20).

[Júnior 2014] Luiz Emygdio Franco da Rosa Júnior. *CORBA v3.0 Specification*. 8ª ed. OMG Document 02-06-33. Renovar, 2014 (citado na pg. 3).

[Malato 2023] Gianluca Malato. *An Introduction to the Shapiro-Wilk Test for Normality*. 2023. url: https://builtin.com/data-science/shapiro-wilk-test (acesso em 21/11/2024) (citado na pg. 21).

[McClenaghan 2024] Elliot McClenaghan. *Mann-Whitney U Test: Assumptions and Example*. 2024. url: https://www.technologynetworks.com/informatics/articles/mann-whitney-u-test-assumptions-and-example-363425 (acesso em 21/11/2024) (citado na pg. 24).

[McKinney 2018] Wes McKinney. *Python para Análise de Dados*. Novatec Editora, 2018 (citado na pg. 19).

[Müller e Guido 2016] Andreas Müller e Sarah Guido. *Introduction to Machine Learning with Python: A Guide for Data Scientists*. 1ª ed. O'Reilly Media, mai. de 2016 (citado na pg. 27).

[UFRN 2021] UFRN. *Tutorial - Regressão Logística*. 2021. url: http://lea.estatistica.ccet.ufrn.br/tutoriais/regressao-logistica.html (acesso em 11/11/2024) (citado na pg. 13).