UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE ESTATÍSTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

Fernando Nepomuceno Valério

MODELO DE RISCO DE FRAUDE EM TRANSAÇÕES DE CARTÃO DE CRÉDITO NO MERCADO DE CRÉDITO BANCÁRIO

Belo Horizonte
2023

Fernando Nepomuceno Valério

MODELO DE RISCO DE FRAUDE EM TRANSAÇÕES DE CARTÃO DE CRÉDITO NO MERCADO DE CRÉDITO BANCÁRIO

Fernando Nepomuceno Valério

Monografia apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de Minas Gerais, como requisito parcial para obtenção do título Especialista em Estatística.

Orientadora: Prof. Sueli Aparecida Mingoti
PhD. em Estatística

Belo Horizonte
2023

Valério, Fernando Nepomuceno.

V164m Modelo de risco de fraude em transações de cartão de crédito no mercado de crédito bancário [recurso eletrônico] / Fernando Nepomuceno Valério - 2023.
1 recurso online (92 f. il., color.) : pdf.

Orientador: Sueli Aparecida Mingoti.

Monografia (especialização) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística.
Referências: f. 78-81.

1. Estatística. 2. Análise de regressão. 3. Regressão logística 4. Cartões de crédito bancário - Europa. 5. Bancos - Fraude. I Mingoti, Sueli Aparecida. II. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística. III. Título.

CDU 519.2(043)

Ficha catalográfica elaborada pela bibliotecária Irénquer Vismeg Lucas Cruz CRB 6/819 - Universidade Federal de Minas Gerais - ICEx

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística
Programa de Pós-Graduação / Especialização
Av. Pres. Antônio Carlos, 6627 - Pampulha
31270-901 – Belo Horizonte – MG

E-mail: pgest@ufmg.br
Tel: 3409-5923 – FAX: 3409-5924

ATA DO 291ª. TRABALHO DE FIM DE CURSO DE ESPECIALIZAÇÃO EM ESTATÍSTICA DE FERNANDO NEPOMUCENO VALÉRIO.

Aos dezenove dias do mês de maio de 2023, às 08:00 horas, com utilização de recursos de videoconferência a distância, reuniram-se os professores abaixo relacionados, formando a Comissão Examinadora homologada pela Comissão do Curso de Especialização em Estatística, para julgar a apresentação do trabalho de fim de curso do aluno **Fernando Nepomuceno Valério**, intitulado: "Modelo de risco de fraude em transações de cartão de crédito no contexto do mercado de crédito bancário", como requisito para obtenção do Grau de Especialista em Estatística. Abrindo a sessão, a Presidente da Comissão, Professora Sueli Aparecida Mingoti – Orientadora, após dar conhecimento aos presentes do teor das normas regulamentares, passou a palavra ao candidato para apresentação de seu trabalho. Seguiu-se a arguição pelos examinadores com a respectiva defesa do candidato. Após a defesa, os membros da banca examinadora reuniram-se sem a presença do candidato e do público, para julgamento e expedição do resultado final. Foi atribuída a seguinte indicação: o candidato foi considerado Aprovado condicional às modificações sugeridas pela banca examinadora no prazo de 30 dias a partir da data de hoje, por unanimidade. O resultado final foi comunicado publicamente ao candidato pela Presidente da Comissão. Nada mais havendo a tratar, a Presidente encerrou a reunião e lavrou a presente Ata, que será assinada por todos os membros participantes da banca examinadora. Belo Horizonte, 19 de maio de 2023.

gov.br Documento assinado digitalmente
SUELI APARECIDA MINGOTI
Data: 23/05/2023 17:02:32-0300
Verifique em https://validar.iti.gov.br

Prof.ª Sueli Aparecida Mingoti (Orientadora)
Departamento de Estatística / ICEX / UFMG

gov.br Documento assinado digitalmente
ELA MERCEDES MEDRANO DE TOSCANO
Data: 24/05/2023 17:19:07-0300
Verifique em https://validar.iti.gov.br

Prof.ª Ela Mercedes Medrano de Toscano
Departamento de Estatística / ICEX / UFMG

Roberto da Costa Quinino:80871291720
Assinado de forma digital por Roberto da Costa Quinino:80871291720
Dados: 2023.05.22 16:51:58 -03'00'

Prof. Roberto da Costa Quinino
Departamento de Estatística / ICEX / UFMG

A Deus, pela minha vida, e por sempre me ajudar a ultrapassar os obstáculos, “a Cristo seja dada a honra, pois dele vem todas as coisas”.

Aos professores pela formação e aprendizado.

A minha esposa Dagmar pelo apoio e companheirismo compreendendo a minha ausência enquanto me dedicava à realização deste trabalho.

RESUMO

Os desdobramentos econômicos, sociais e culturais pelos quais a sociedade passou, em especial nas últimas décadas, foram fatores preponderantes para a expansão da utilização do cartão de crédito como principal meio de pagamento.

A revolução tecnológica, inseriu milhões de novos usuários no mundo digital e esses usuários passaram a utilizar equipamentos de comunicação, permitindo, assim, a realização de transações de compra e venda e a criação de um novo mercado digital.

Dessa forma, a criação de novos mecanismos de segurança para evitar perdas com fraudes tornou-se um desafio para as instituições bancárias, uma vez que as relações entre os bancos e clientes passaram a ser cada vez mais digitais.

O presente trabalho faz um breve relato do desenvolvimento do mercado de cartão de crédito, descrevendo as principais formas de fraude. Utiliza-se uma abordagem para prever transações legítimas ou fraudulentas, através de um modelo de regressão logística, no conjunto de dados de detecção de fraude no cartão de crédito da Kaggle.

O trabalho utilizou uma base de dados que contém transações feitas por cartões de crédito em setembro de 2013 por titulares de cartões europeu, portanto, as conclusões não podem ser estendidas para o cenário brasileiro, assim qualquer implementação no cenário brasileiro deve ser adaptada de acordo com as características e necessidades específicas.

Propõe-se duas formas distintas de tratamento dos dados: a primeira, utilizando a remoção de variáveis correlacionadas e a segunda, realizando a transformação dos dados através da Análise de Componentes Principais (PCA). O modelo de regressão logística é aplicado em ambas as formas de tratamento de dados, avaliando-se o seu desempenho.

Ao final, os resultados demonstram que o modelo com aplicação da Análise de Componentes Principais apresenta melhor resultado para o estudo em questão.

Palavras-chave: **mercado de cartões de crédito**; **fraudes bancárias; regressão logística; análise de componentes principais.**

ABSTRACT

The economic, social and cultural developments that Brazilian society has gone through, especially in recent decades, were preponderant factors for the expansion of the use of credit cards as a means of payment, making them the most used today. Another important factor was the technological revolution, which introduced millions of new users to the digital world. These users began to use communication equipment, thus allowing purchase and sale transactions and the creation of a new digital market.

In this way, the creation of new security mechanisms to avoid losses due to fraud has become a challenge for banking institutions, since the relationships between banks and customers have become increasingly digital.

This monography makes a brief account of the development of the Brazilian credit card market, describing the main forms of fraud.

A logistic regression model was applied using Kaggle's credit card fraud detection dataset with the purpose of predicting legitimate or fraudulent transactions.

Two different ways of treating the data was proposed: the first, using the removal of correlated explanatory variables and the second, transforming the data through Principal Components Analysis, also known as PCA. The logistic regression model is applied to both forms of previous data treatment and their performance were evaluated.

The model with the application of Principal Components Analysis presented better results for the study in question.

This work is divided into ten parts, which are introduction, contextualization, problem definition, justification, objectives, data origin, data sampling, descriptive analysis, predictive model and final considerations.

Keywords: **credit card market; bank fraud; logistic regression; principal component analysis.**

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

BACEN – Banco Central do Brasil

TED – Transferência Eletrônica Disponível

PIX – Pagamento Instantâneo

PIB – Produto Interno Bruto

BNDES – Banco Nacional de Desenvolvimento Econômico e Social

FEBRABAN – Federação Brasileira dos Bancos

FEEB-PR - Federação dos Empregados em Estabelecimentos Bancários no Estado do Paraná

CERT - Centro de Estudos, Resposta e Tratamento de Incidentes de Segurança no Brasil

IGP-DI - Índice Geral de Preços – Disponibilidade Interna

LISTA DE GRÁFICOS

SUMÁRIO

1. INTRODUÇÃO

O Brasil vem passando por enormes transformações nas últimas décadas, em especial no se diz respeito as relações comerciais e como as pessoas interagem entre si e com as empresas. Estas mudanças são reflexo da globalização mundial que cada vez mais conecta pessoas ao mundo, reduzindo barreiras e permitindo que pessoas mesmo que fisicamente distantes possam se comunicar e se relacionar.

Dentre as transformações ocorridas, uma das mais importantes, principalmente no contexto econômico, foi a expansão do crédito associada a ampliação da utilização do cartão como meio de pagamento.

As décadas de 1990 e 2000 foram importantes para o desenvolvimento e a consolidação do cartão como principal meio de pagamento da população brasileira e, tais transformações, serão citadas nas seções seguintes.

A agência pública de notícias divulgou que segundo dados do Banco Central do Brasil já em 2021 as transações por cartão foram as maiores em volume de transações totais realizadas no Brasil com 51,1% seguida do PIX com 16,2% débito direto (11,4%), boleto (9,8%), convênios (5,1%), TED (2,1%), transferências interbancárias (1,8%), e outros (2,5%).

Reportou ainda que em relação ao canal de pagamento, o celular foi responsável por 60% da quantidade de transações, seguido das realizadas na internet *banking* (17,8%); correspondentes bancários, como lotéricas (13,3%); agências de atendimento (5,7%); caixas 24 horas (2,4%), e outros (0,8%) **(EBC, 2022)**.

Com o crescimento do mercado de cartões, também cresceu o volume de eventos de quebra de segurança e risco de fraudes e golpes. Simultaneamente, aprimoraram-se os golpes que passaram a envolver mais conhecimento tecnológico e menos ações presenciais.

Parte dessa evolução acontece porque hoje, na maior parte das vezes, para que um fraudador aplique um golpe ou invada um sistema, ele não precisa estar presente no local que sofrerá o dano, o que acaba fazendo com que o anonimato seja um incentivador para a prática do crime.

Nesse contexto, as instituições financeiras, em especial os bancos, tem um papel primordial por serem os responsáveis pela concessão de crédito e definição

de risco associado a esse crédito. Dessa forma, torna-se cada vez mais importante a criação de mecanismos de autenticação de segurança para tais transações, uma vez que, na maioria, essas transações ocorrem digitalmente.

Diante de tal cenário, o presente trabalho visa dar sua contribuição no sentido de contextualizar os principais fatos históricos que permitiram o crescimento do mercado de cartões. Simultaneamente, contextualizar a evolução dos mecanismos de fraude nesse período.

2. CONTEXTUALIZAÇÃO

2.1. O Mercado de Cartão de Crédito no Brasil

A década de 1990 foi marcada por enormes mudanças nas esferas política, econômica e social do Brasil, trazendo alterações nas relações pessoais e institucionais. Após um longo período de protecionismo da economia, temos em 1990 o rompimento desse ciclo por meio da abertura comercial e da busca da ampliação do mercado brasileiro através do comércio internacional. O aumento do volume do comércio global, associado à revolução nos meios de comunicação, em especial nas telecomunicações e no crescimento de novos equipamentos, impulsionaram decisivamente o processo de abertura do comércio brasileiro. Conforme descrito por **SANTOS (2009)**:

> Esses anos marcam a consolidação de uma nova era da globalização e do desenvolvimento tecnológico, influenciando empresas, governos, a opinião pública, entre outros **(SANTOS, 2009).**

Tais mudanças fizeram com que o país desse os primeiros passos em direção a um novo modelo pautado pela busca do livre mercado, a redução da participação do Estado na economia e a privatização de estatais.

Outro fator extremamente importante para a consolidação dos fundamentos que permitiram a ampliação do mercado de cartões de crédito, sem dúvida, foi a estabilidade da moeda a partir da criação do Plano Real.

O Plano Real foi implementado em 1994 no governo Itamar Franco, que combateu a hiperinflação que vigorava no Brasil desde 1979. Esse cenário foi causado por políticas econômicas equivocadas praticadas no período de 1986 a

1991, quando houve 5 (cinco) mudanças de planos econômicos: Cruzado Novo – 1986; Plano Bresser – 1987; Plano Verão – 1989; Plano Collor I – 1990; Plano Collor II – 1991. Tais mudanças provocaram diversas modificações dos vetores para cálculo de índices de correção monetária, controle de preços e congelamentos de ativos financeiros, confisco de correção monetária, dentre outros **(CYSNE e COSTA, 1996)**.

O Plano Real foi ancorado em 3 (três) pilares estruturais: a âncora cambial, a abertura econômica e a base monetária rígida (juros altos), conforme descrito por **SILVA (2015)**.

A âncora cambial se pautava na ideia de criar uma paridade entre a nova moeda Real e as moedas estrangeiras, em especial o Dólar, com a intenção de controlar a cotação de uma moeda em relação a outra. Essa medida buscava reduzir os impactos dos efeitos externos sobre a economia brasileira e também reduzir o custo da dívida que, em sua maior parte, estava lastreada em dólar. A cotação inicial da moeda teve ajuste a partir dos movimentos de saída de capitais e outros movimentos que faziam com que o Banco Central atuasse através do aumento de juros.

CYSNE e COSTA (1996) detalham os efeitos desses movimentos de ajuste bem como suas consequências:

> Em adição, o crescimento residual dos preços (principalmente dos serviços) implicou um resíduo cada vez menor do salário para saldar as prestações. O resultado foi o aumento da inadimplência e todas as suas desagradáveis consequências sobre a saúde do sistema financeiro **(CYSNE e COSTA, 1996)**

O movimento de abertura econômica seguia seu curso através das privatizações e da abertura das relações comerciais com países da América Latina e Europa, aliada à política de altos juros que buscava evitar a evasão de divisas.

As mudanças realizadas pela implantação do Plano Real representaram um momento extremamente disruptivo para o setor bancário, que sofreu um grande impacto em seu modelo de remuneração, antes ancorado nas transferências inflacionárias.

Assim, as instituições bancárias passam a ter perdas de rentabilidade de seus produtos (40% em mês para uma média de 3,65% IGP-DI). Nos 4 (quatro) anos anteriores da implantação do Plano Real, os juros pagos por depósito à vista, menos

os encaixes totais, giraram em torno de US$ 794,784 milhões ao mês. Já no período de junho1994 a julho 1995 estes valores foram à ordem de US$ 8.631 milhões **(CYSNE e COSTA,1996)** o que obrigou a uma mudança onde a busca por novos meios de rentabilização foi o ponto-chave.

Os anos que se seguiram (2003 a 2010) foram anos de significativo aumento do crédito pessoal, devido a um conjunto de políticas de incentivo promovidas pelo governo com intuito de fomentar a expansão de crédito e fortalecer o sistema financeiro.

O crédito aumentou expressiva e continuamente entre os anos de 2003 a 2010, inclusive após a crise de 2008. Assim, o volume de crédito, que representava 26% do produto interno bruto em dezembro de 2002, atingiu 45,2% do PIB em dezembro de 2010 **(MORA, 2015).**

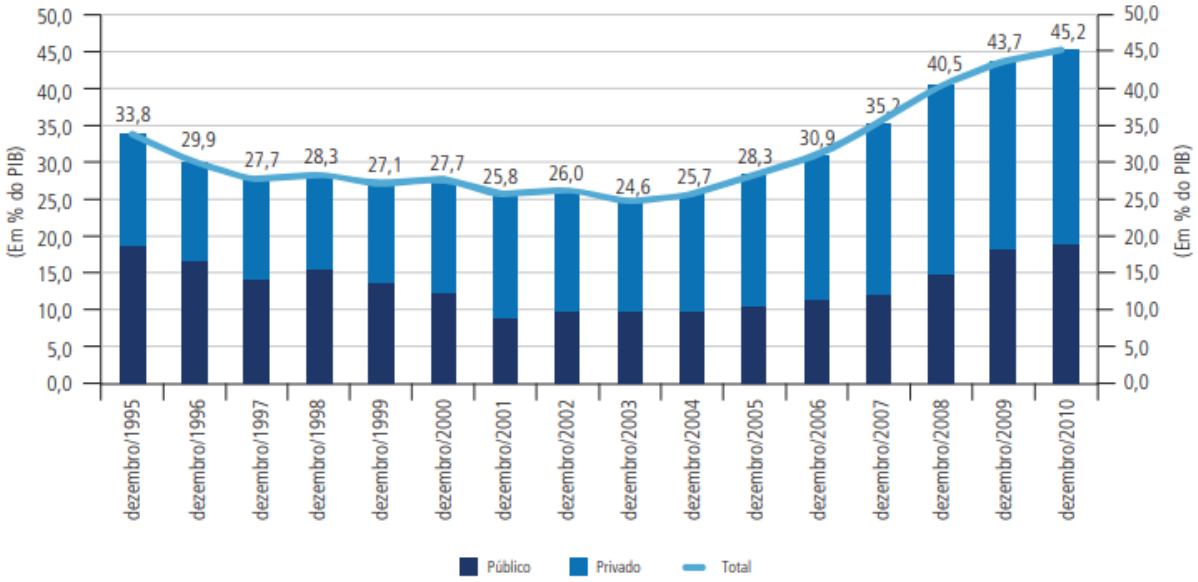
Essa elevação do volume do crédito em um contexto macroeconômico caracterizado por elevadas taxas de juros, foi inicialmente capitaneada pelos bancos privados e ocorreu tanto no âmbito da pessoa física quanto da jurídica.

Quando comparado a outros países, o volume de crédito na economia brasileira era relativamente baixo e, mesmo com o crescimento observado entre 2003 e 2008-2009, manteve-se abaixo de países como China, Chile e Austrália **(MORA, 2015).**

Os bancos privados foram os principais agentes da ampliação do crédito pessoal nesse período, muito impulsionados pela alta da taxa básica de juros que atuava como uma alavanca para maiores ganhos.

A criação do crédito consignado em 2003 permitiu aos trabalhadores vinculados a determinados sindicatos e aos servidores públicos e aposentados o acesso ao crédito bancário a taxas de juros proporcionalmente mais baixas que já representavam 3,7p.p. do PIB em dezembro de 2010, e explicavam o comportamento do crédito pessoal **(MORA, 2015).**

Gráfico 1 - Evolução do crédito por controle de capital (1995-2010)



Fonte: MORA 2015.

Conforme demonstrado no Gráfico 1, o ciclo de 2003 a 2010 foi marcado por forte crescimento do crédito pessoal tanto para pessoas físicas quanto para pessoas jurídicas devido à estabilidade econômica brasileira e ao crescimento.

Nesse período houve uma expansão significativa de políticas de incentivo à concessão de crédito como, por exemplo, a Lei Federal n. º 10.820/2003 que rege sobre as diretrizes para concessão de empréstimos consignados, que buscava trazer maior segurança aos bancos e, com isso, reduzir os juros do crédito ao consumidor. Também houve medidas de ampliação de crédito às empresas por meio dos bancos com participação do governo (Banco do Brasil e Caixa Econômica), bem como por meio do BNDES (Banco Nacional de Desenvolvimento Social e Econômico).

Essa expansão só foi possível devido aos pilares macroeconômicos estabelecidos pelo governo anterior que permitiu um maior controle dos gastos públicos, o controle da inflação por meio de medidas monetárias restritivas e a manutenção do valor da moeda por meio do câmbio flutuante. Tal política conhecida como “O Tripé Macroeconômico” permitiu estabelecer o melhor funcionamento da economia brasileira que, no passado, tanto sofreu com problemas de inflação, gastos públicos descontrolados e enfraquecimento das reservas internacionais.

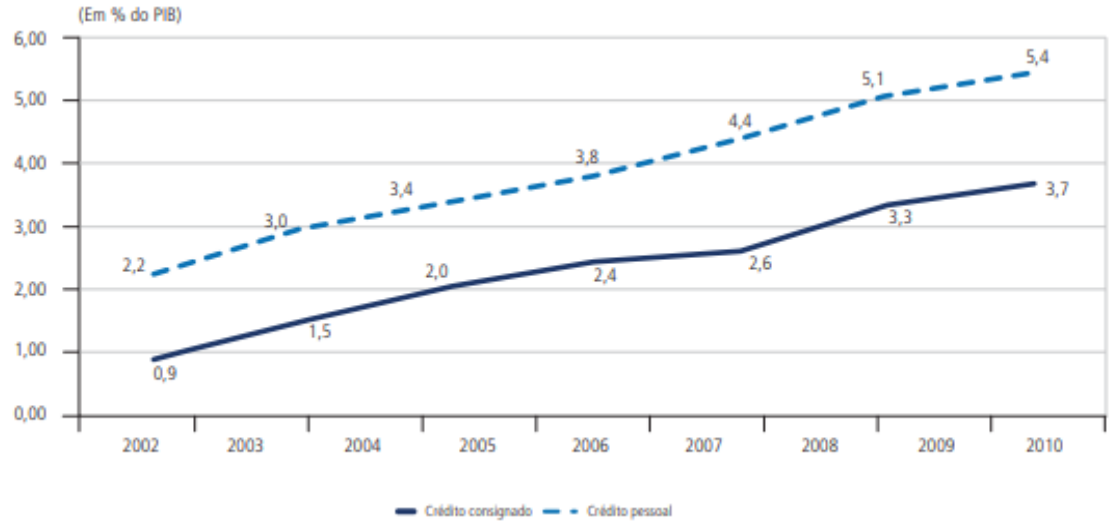
Esse movimento resultou na ampliação da base de crédito do país, mas, gerou uma maior instabilidade no mercado financeiro, ocasionando a elevação das

taxas de crédito ao consumidor. Esta elevação acontecia, pois, ao mesmo tempo em que se aumentava a base de crédito, concomitantemente crescia o índice de inadimplência, aumentando assim o risco do crédito concedido.

A lei que permitia a concessão de crédito consignado buscava reduzir as incertezas quanto ao pagamento dos empréstimos, uma vez que esse era debitado diretamente da folha de pagamento do funcionário, trazendo assim maior garantia de pagamento.

> A introdução do consignado alterou o perfil do crédito pessoal ao repercutir tanto sobre a taxa de juros média ao ano (a.a.) (que descendeu de um patamar de 80% a.a. para aproximadamente 40% a.a.),8 quanto sobre o prazo (com a ampliação do prazo médio de duzentos dias para mais de 550 dias) (Gráfico 2). Ou seja, a possibilidade de consignação em folha de pagamento ocasionou a redução do custo dos empréstimos caracterizados como crédito pessoal à pessoa física e, simultaneamente, permitiu um aumento do prazo. Este processo possibilitou uma redução expressiva do valor das prestações e, portanto, do comprometimento da renda dos tomadores de crédito **(MORA, 2015, p.16).**

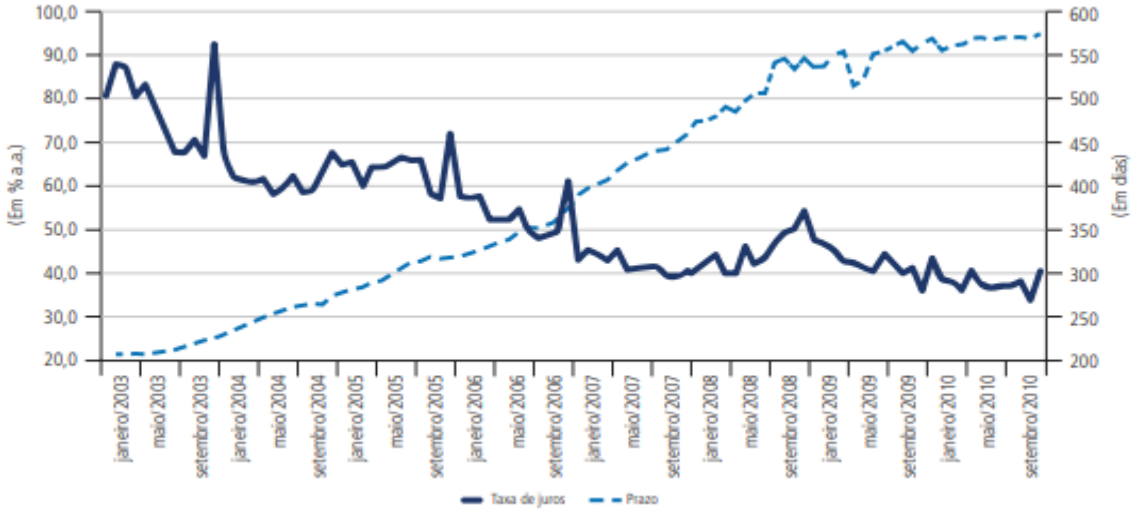
Gráfico2- Evolução do crédito pessoal e do consignado (dez./2004-dez/2010)



Fonte: Bacen -series temporais apud (**MORA 2015 p.16**).

No período de 2002 a 2010 o crédito consignado e crédito pessoal tiveram um aumento de mais de 100% em relação a participação do PIB.

Gráfico 3- Evolução das operações de crédito: crédito pessoal, taxa de juros média a.a. e prazo médio em dia (2003-2010)



Fonte: BACEN -series temporais apud (**MORA 2015 p.17**).

De acordo com o Gráfico 3, observa-se dois movimentos que foram grandes impulsionadores da ampliação do crédito: a redução das taxas de juros dos empréstimos e a ampliação do tempo médio dos empréstimos.

2.2. Evolução do Cenário de Fraudes no mercado Brasileiro

O mercado mundial vem passando significativas mudanças nos meios de transação e comunicação entre as pessoas e empresas. Estas mudanças são reflexos de uma grande e importante transformação da sociedade que ocorre a partir de uma revolução tecnológica, promovendo novos marcos políticos, econômicos, culturais e morais, e fazendo com que as pessoas se tornem cada vez mais conectadas às novas tecnologias.

Os novos meios de comunicação e a informatização permitem que pessoas pelo mundo todo possam, através de seus dispositivos eletrônicos, realizar operações de compra, venda, comunicações, etc. com velocidades nunca imaginadas, o que faz com que as instituições busquem novos mecanismos de segurança e proteção. Para isto, torna-se cada vez mais necessária a garantia da segurança e confiabilidade dos dados, bem como da segurança de autenticidade, identificação do originador, identificação do receptor, garantia de que a transação e legal e está sendo feita com a livre vontade das partes.

Nesse ponto, um dos grandes desafios vividos pelas instituições é o de evitar as transações de fraude, que tem como definição o “ato ou dito com intenção de enganar ou prejudicar alguém, ação ilícita, punível por lei, que visa enganar alguém ou alguma entidade ou escapar de obrigações legais” **(PRIPERAM, 2022).**

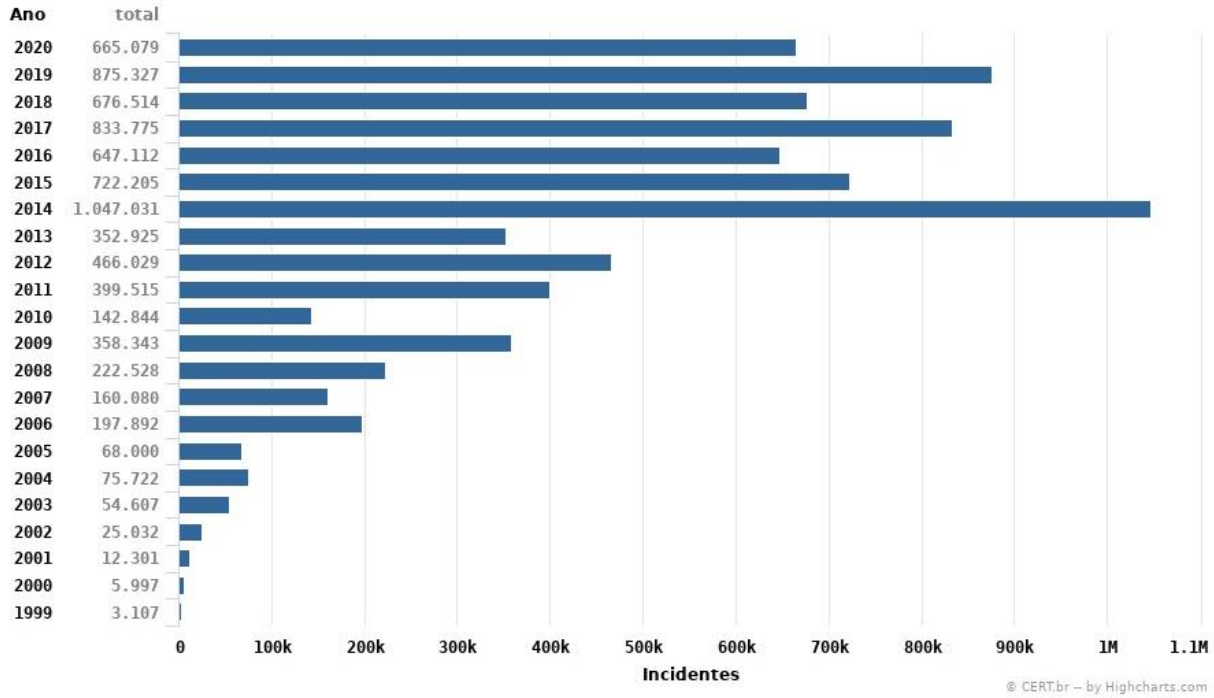
Esses se manifestam de diversas formas, sendo que para alguns existem previsões legais para punição e para outros ainda não. **BASTOS e PEREIRA (2007)** discorrem sobre isto dizendo:

> De tempo em tempo surgem novos tipos de delitos, os quais talvez por ainda não encontrarem cominação legal, não o podem ser assim chamados, senão como condutas imorais, antiéticas e lesivas. Para citar alguns, independente da previsão em lei, temos: fraudes cometidas mediante manipulação de computadores, falsificações informáticas de dados e documentos, danos ou modificações de programas ou dados computadorizados (vírus, acesso não autorizado de hackers, etc.), espionagem industrial, sabotagem de sistemas, pesca ou averiguação de senhas, pornografia infantil, jogos de azar e lavagem de dinheiro **(Furlaneto Neto / Guimarães, 2003, p. 70-71 *apud* BASTOS e PEREIRA, 2007, p.3).**

O Centro de Estudos, Resposta e Tratamento de Incidentes de Segurança no Brasil (**CERT, 2022**) é uma entidade mantida pela NIC.br responsável pela operação do domínio “.br”, bem como pela distribuição de números de IP e do registro de sistemas autônomos no Brasil.

Essa Instituição registra os reportes de incidente relacionados a crimes cibernéticos no país como apresentado no Gráfico 4.

Gráfico 4- Total de Incidentes Reportados ao CERT.br por Ano. (1999 a 2020)



Fonte: **CERT, 2022**– Incidentes.

No Gráfico 4 verifica-se a evolução do volume de registros de fraudes eletrônicas com destaques para os anos de 2005 e 2006 com aumento de 191%, período em que houve ampliação do crédito pessoal no Brasil. Nesse período, o destaque dos tipos de ataques foram o Worm-55%(atividade maliciosa relacionada com o processo automatizado de propagação de códigos maliciosos na rede), Scan-23% (escaneamento de serviços ativos no computador com intuito de identificar possíveis vulnerabilidades) e Fraude-21% (qualquer ato ardiloso, enganoso, de má-fé, com intuito de lesar ou ludibriar outrem, ou de não cumprir determinado dever).

De 2006 a 2013 temos um crescimento de mais 78,3%, demonstrando uma ascendência constante dos relatos, com destaque para registros de Scan - 46,66% e Fraude – 24,28%.

A partir de 2014 vê-se um crescimento exponencial dos dados de fraude, com destaque para registros a partir de setembro de 2014 onde houve estouro do volume de casos de Fraude – 44,66% e Scan – 25,18% e DoS – 21,39% (*Denial of Service – o atacante utiliza um computador ou conjunto de computadores para tirar de operação um serviço, servidor ou rede*).

Em 2020 os destaques continuam com Scan – 59,85%, Worm – 20,15%, e DoS – 10,25% sendo que a fraude teve apenas 4,66%.

O cenário de 2020 mostra claramente o quanto os ataques têm se tornado cada vez mais especializados, demonstrando maior grau de especificidade e conhecimento por parte dos fraudadores, e não só isso, mas a cada dia nova forma de quebra de segurança e novos mecanismos com intuito de lesar a outros.

2.3. Fraude em cartão de Crédito

Quando se trata especificamente de fraude em cartão de crédito, além dos cenários já descritos, temos outras modalidades utilizadas pelos malfeitores como:

1. Perda ou roubo do cartão: nesta modalidade o cartão em posse do cliente é roubado ou o cliente perde o mesmo e o fraudador busca utilizar o cartão gastando o maior valor possível no menor prazo antes que o cliente registre a perda/roubo junto a sua operadora de cartão de crédito;
2. Extravio de cartão: acontece quanto o cartão é interceptado ou por outro motivo não é entregue ao seu dono legítimo. Nesta modalidade o cartão é desviado entre o momento de sua emissão e o envio ao cliente final;
3. Fraude de proposta: nesta modalidade o fraudador de posse dos dados do cliente se passa pelo mesmo buscando aprovar uma proposta de novo cartão de crédito em nome do verdadeiro dono dos documentos;
4. Invasão da conta: acontece quando um terceiro se utilizando de conhecimento especializado invade a conta do cliente e realiza transações em nome do mesmo sem seu consentimento.
5. Engenharia Social: o fraudador se passando pela instituição do cartão convence o cliente a informar dados sigilosos e com estes invade a conta;
6. Phishing: utiliza-se de um meio de comunicação eletrônico, como e-mail e/ou WhatsApp e convence o cliente a clicar em determinado link que o leva à exposição dos dados ou, ainda, a transferir determinada quantia, se passando por outra pessoa;
7. Malware: também conhecido como “software malicioso”, assim como o Phishing esse é encaminhado para o cliente por meio de algum meio

eletrônico, normalmente e-mail, e quando o cliente clica em um determinado link o software malicioso entra em ação e rouba os dados do cliente;

8. Falsificação manual: o fraudador, em um momento de compra direta do cliente em um estabelecimento, aproveita de um momento de descuido do cliente e realiza a cópia direta dos dados do cartão, bem como os dados do código de segurança e os utiliza posteriormente;
9. Falsificação eletrônica: o fraudador adquiri por meios escusos os dados dos cartões de clientes e os utiliza na tentativa de realizar compras por meio digital;
10. Mail Order: o fraudador envia e-mail para o cliente se passando pelo funcionário da instituição (banco), informando a necessidade de recolhimento de cartão supostamente inválido. Após o recolhimento, utiliza os dados desse cartão para realização da fraude, uma vez que o mesmo não está cancelado na instituição de origem.
11. Telefone Order (Moto): um motoboy se apresenta ao cliente como representante de alguma empresa reconhecida de mercado e informa sobre um brinde que o cliente ganhou. Então, obtém os dados do cliente, juntamente com um procedimento de validação do cartão. Nesse momento, a máquina utilizada pelo fraudador captura os dados do cartão do cliente utilizados para realização da fraude.
12. Laranjas: são casos onde o fraudador alicia pessoas a cederem seus dados pessoais, utilizando-os para as fraudes.

O processo de segurança antifraude se divide em etapas de proteção que possuem suas características específicas conforme o momento de segurança a ser aplicada. A quantidade de etapas pode variar de instituição para instituição ou ter subclassificações, mas de forma geral se distribuem da seguinte forma:

1. Prevenção: essa etapa acontece no momento da aprovação da transação do cliente, quando se tem a aplicação de regras de segurança que permitem avaliar padrões de comportamento para criação de regras que impeçam a fraude. Tem-se, ainda, a aplicação de Modelos Estatísticos, Machine Learning, Inteligência Artificial, dentre outros. Como nessa etapa a transação do cliente ainda não foi concluída, ela é extremamente sensível; por isso,

tende a ter um nível de assertividade e acurácia extremamente elevados, para evitar erros na experiência de compra do cliente;

2. Detecção: essa etapa acontece após a aprovação da transação do cliente. Nesse momento, são realizadas avaliações de risco após compra e, em caso de algum problema, são tomadas ações necessárias para o restabelecimento da segurança, e correção do incidente. Também são realizadas análises de padrões, avaliação de pós-compra, ações de contingência de ataques, dentre outros. Estas ações são extremante importantes para contenção de problemas e para atuação em casos de ataques.
3. Contestação (desconhecimento de transação): acontece quando as ferramentas de segurança não conseguiram identificar o problema na transação, e o cliente realizou o acionamento diretamente na instituição, informando do desconhecimento da transação. Passa-se a realização de avaliações quanto às alegações do cliente para apurar a procedência ou improcedência. Após determinação da etapa anterior, são realizadas análises na de melhorias em regras e modelos.

3. DEFINIÇÃO DO PROBLEMA

O setor financeiro mudou drasticamente nas últimas décadas com o advento da transformação digital e das novas realidades da economia global. Serviços financeiros diários, soluções de investimento, sistema de bolsa de valores, soluções corporativas, tudo isso foi digitalizado e agora vive-se uma nova realidade. Tudo isso fez com que surgissem novos players no mercado de serviço financeiro, as chamadas Fintechs e os Bancos Digitais.

Os Bancos Digitais são empresas regulamentadas pelo Banco Central e com autorização para atuarem como banco, oferecendo produtos e serviços tais como: investimentos, transferências, pagamentos, conta corrente, cartão de crédito, contratação de seguros, solicitação de empréstimo online, dentre outros, assim como os bancos tradicionais.

As Fintechs, por sua vez, são empresas que fornecem serviços financeiros com as mesmas características das instituições financeiras tradicionais (bancos, corretoras, seguradoras, gestoras de fundos, casas de câmbio), e que possuem a

inovação tecnológica na base das suas operações, assim como os Bancos Digitais. Contudo, tais instituições possuem regulamentação mais simples que as dos Bancos Digitais, com menor necessidade de incorporação de investimentos e uma regulamentação menos complexa.

Nessa nova era digital, quando instituições financeiras tradicionais, Fintechs e Bancos Digitais convivem e disputam a atenção e as transações financeiras da população brasileira, o desenvolvimento tecnológico tem sido o grande diferencial no crescimento do setor, o que, por sua vez, produz uma gama de soluções digitais que visam facilitar a relacionamento com os clientes, desde o atendimento simplificado até os tipos de forma de investir cada vez mais customizado.

O BACEN (Banco Central do Brasil) descreve esse movimento como algo que engloba todas as instituições financeiras conforme vemos a seguir:

> O processo de digitalização dos serviços bancários surgiu da necessidade de desburocratização dos processos dos grandes bancos, o que resultou no aprimoramento da experiência do cliente, que teve acesso a mais segurança, transparência e agilidade em suas operações. Para lidar com esse movimento de digitalização bancária, os bancos têm buscado o desenvolvimento de capacidades organizacionais internas ou parcerias com empresas de tecnologia. O fenômeno da digitalização bancária engloba, em menor ou maior grau, todas as instituições financeiras, incluindo os maiores conglomerados bancários, que já estão em processo de migração integral para o ambiente digital e têm estimulado a migração dos clientes para esse canal de atendimento **(BACEN, 2019).**

Ainda segundo a FEBRABAN (Federação Brasileira de Bancos), as transações financeiras realizadas por meio de dispositivos eletrônicos cresceram 11% em 2019, registrando 89,9 bilhões de operações, representando 44% do total de transações financeiras.

> “O mobile banking transformou-se em uma expressiva porta de entrada para a inclusão financeira de milhões de brasileiros pela possibilidade de carregar no bolso e acessar, em qualquer hora ou local, serviços antes restritos a agências bancárias”, afirma Gustavo Fosse, diretor setorial de Tecnologia e Automação Bancária da **FEBRABAN (2020**).

Essa expansão do móbile desencadeou uma disparada na criação de aplicativos de alta qualidade para as instituições bancárias mais tradicionais. Todas elas já tinham protótipos de aplicativos, mas ainda enxergavam a onda mobile como algo distante do público brasileiro.

Nesse novo cenário, a FEEB-PR (Federação dos Empregados em Estabelecimentos Bancários no Estado do Paraná) diz:

> As instituições financeiras exclusivamente digitais chegaram com diferenciais agressivos para a época, como cartão de crédito sem anuidade e conta corrente totalmente gratuita. Ainda que não fosse uma inovação de fato, pois bancos digitais com essa proposta já existiam em outros países, essa abordagem representou uma gigantesca [quebra de paradigma] para o setor bancário brasileiro **(FEEB-PR, 2021).**

Apesar do grande crescimento desse novo modelo associado aos bancos digitais, muitos são os desafios enfrentados por essas instituições como, por exemplo: a grande necessidade de investimentos, a busca por novos produtos atrativos, a necessidade contínua de desenvolver e ampliar a segurança das transações, a melhoria dos mecanismos de comunicação com o cliente e a maior exposição de risco financeiro devido a margens de ganho menos expressivas em comparação aos bancos tradicionais.

Os bancos digitais são regulamentados pelas mesmas regras dos bancos tradicionais, não possuindo regulação própria. Tal cenário se dá devido à dificuldade de se regulamentar regras para um mercado que tem nas mudanças e na inovação como cerne de sua natureza.

> Entre os documentos que regulam o funcionamento dos bancos digitais, há algumas que valem também para os bancos tradicionais, mas para as suas modalidades de contas digitais. Como é o caso da Resolução CMN nº 3.919, de 2010, que "altera e consolida as normas sobre cobrança de tarifas pela prestação de serviços por parte das instituições financeiras e demais instituições autorizadas a funcionar pelo Banco Central do Brasil e dá outras providências" (Conselho Monetário Nacional[CMN], 2010, p. 1). Essa resolução foi um importante passo por considerar como um serviço bancário essencial e sobre o qual é vedada a cobrança de tarifas a "prestação de qualquer serviço por meios eletrônicos, no caso de contas cujos contratos prevejam utilizar exclusivamente meios eletrônicos " **(CMN, 2010, p. 2 apud MACIEL, 2018, p.5).**

Uma vez que abertura da conta digital acontece de forma não presencial, a segurança aparece entre as principais preocupações do BACEN em relação à utilização do meio eletrônico para tal.

> As tecnologias atualmente disponíveis possibilitam que as instituições estruturem seus controles internos de modo a garantir um processo de abertura de conta seguro e eficaz, propiciando maior comodidade a seus clientes. Nesse sentido, é possível implementar mecanismos de segurança

> para verificar as informações relativas aos clientes, entre as quais se destacam a validação de dispositivos eletrônicos de acesso aos sistemas da instituição, a geolocalização do cliente, a captura de cópia de documentos e foto digitais, gravação de imagem e detecção de presença do usuário, assinatura digital etc., mantendo todos os dados necessários ao acompanhamento do processo e ao rastreamento das informações para fins de auditoria **(BACEN, 2016c, p. 66 apud MACIEL, 2018, p.7).**

O BACEN ainda regulamenta sobre o processo de digitalização de documentos e de manutenção de documentos, digitalizados definindo que estes devem ter:

> I - integridade, autenticidade, confidencialidade E possibilidade de rastreamento do documento digitalizado;
> II - proteção do documento digitalizado contra o acesso, o uso, a alteração, a reprodução e a destruição não autorizados;
> III - rastreamento e auditoria dos procedimentos empregados;
> IV - padrão de qualidade da imagem do documento digitalizado que garanta a sua legibilidade e uso; e
> V - indexação que possibilite a localização, o gerenciamento e a preservação do documento digitalizado, bem como posterior conferência da regularidade das etapas do processo adotado **(CMN, 2016a, p. 5 apud MACIEL, 2018, p.6).**

E, por fim, trata de forma mais específica sobre a política de segurança cibernética dos dados e das transações realizadas através da resolução CMN n. º 4.658 que diz:

> A Resolução CMN nº 4.658, trata da obrigatoriedade da implementação e da manutenção de uma política de segurança cibernética e sobre os requisitos para a contratação de serviços de processamento e armazenamento de dados e de computação em nuvem (CMN, 2018b). A resolução considera "a crescente utilização de meios eletrônicos e de inovações tecnológicas no setor financeiro, o que requer que as instituições tenham controles e sistemas cada vez mais robustos, especialmente quanto à resiliência a ataques cibernéticos" (BACEN, 2018, p. 1). Por abordar a segurança cibernética, essa resolução não tem efeito apenas sobre os bancos digitais, mas sobre as demais instituições também. Para D'Andrea (2018, p. 1), essa resolução permite que "que as instituições reguladas avancem de maneira estruturada em um mundo cada vez mais digital, melhorem a relação de confiança com o mercado e sejam efetivas na gestão de riscos, no compliance e controles internos, enfim, na governança cibernética" **(MACIEL, 2018, p.7).**

Desta forma, tantos bancos digitais e quanto tradicionais têm investido cada vez mais em segurança e em conhecimento a partir da análise de dados para garantir maior integridade e confiança ao sistema e garantir a confiabilidade nas transações digitais.

4. JUSTIFICATIVA

O cenário de segurança transacional nas instituições bancárias tem se tornado cada vez mais desafiador. É grande a necessidade de investimentos, a busca por novos produtos atrativos, a necessidade contínua de desenvolver e ampliar a segurança das transações e a melhoria dos mecanismos de comunicação com o cliente.

Dessa forma, os bancos têm investido cada vez mais em tecnologia para segurança de suas transações, buscando a identificação dos agentes envolvidos nas transações.

Além disso, as instituições bancárias são rigorosamente regulamentadas pelo BACEN no sentido de garantirem integridade, confiabilidade, segurança e identificação das transações ocorridas em sua estrutura transacional.

5. OBJETIVO

5.1. Objetivo Geral

Contextualizar o mercado de cartão de crédito no Brasil e propor através da aplicação de técnicas de estatística multivariada um modelo de detecção de fraudes em transações de cartão de crédito.

5.2. Objetivo Específico

- Verificar o desenvolvimento do mercado brasileiro de cartões de crédito.
- Construir um modelo de detecção de fraudes em transações de cartão de crédito.

6. ORIGEM DOS DADOS

Em todo trabalho de modelagem estatística, uma das etapas mais importantes é a análise descritiva dos dados. Nela, podemos analisar as principais características dos dados, identificando padrões e definindo o nível inicial de importância das variáveis envolvidas no modelo, além de permitir ao pesquisador tomar a melhor decisão sobre a melhor forma de utilizar os dados.

Nesta monografia, será utilizada como auxílio uma base de dados pública disponibilizada pela Kaggle (https://www.kaggle.com/), que é a maior comunidade na internet relacionada ao aprendizado de ciência de dados, com mais de 536 mil membros ativos. As características da base, bem como os detalhes de sua construção, foram originalmente extraídas durante uma colaboração de pesquisa entre a Worldline e o Machine Learning Group (http://mlg.ulb.ac.be) da ULB (Université Libre de Bruxelles) sobre mineração de big data e detecção de fraudes. O trabalho original publicado em 2021 possuía as seguintes características:

> O banco de dados de origem é composto por cerca de 143 milhões de transações de comércio eletrônico que ocorreram no país de origem durante 183 dias (91 de treinamento, 10 de validação e 82 de teste). A taxa de fraude na origem é de 0,13% e cada transação é descrita por 23 características (em particular, não há dados geográficos sobre os titulares de cartão, pois o país é diferente para ambos os domínios). O banco de dados de destino é composto por cerca de 60 milhões de transações de comércio eletrônico do país de destino (os mesmos dias e recursos do banco de dados de origem) com uma taxa de fraude de 0,21% **(LEBICHOT, 2021).**

No trabalho de **Lebichot (2021)** foram utilizadas diversas técnicas de tratamento dos dados das quais destaca-se duas: o Treinamento de Classificador e a posteriormente Análise de Componentes Principais (PCA) conforme descrito:

> Treinamos um classificador (por exemplo, DNN ou EE) no conjunto de dados de origem. O classificador é então usado para retornar uma estimativa da probabilidade condicional para cada fonte e amostra de destino. Esses valores são usados para criar um recurso adicional *Pred1* para aumentar o conjunto de dados original.
>
> Como resultado, aumentamos o conjunto de dados original com vários novos recursos: *Pred1* e Pca1, Pca2, … O número de componentes PCA, nPC, é ajustado durante a validação. A expectativa é que tais características possam codificar no conjunto de treinamento o relacionamento entre as distribuições fonte e alvo, tanto de uma perspectiva marginal quanto condicionalmente dependente.
>
> Construímos uma análise de componentes principais (por exemplo, PCA) no conjunto de treinamento de origem. As projeções de todas as transações (origem e destino) nos primeiros PCs retornam várias variáveis adicionais

> (denotadas *PCA*) para aumentar o conjunto de dados original **(LEBICHOT, 2021).**

A técnica DNN/EE é uma abordagem que utiliza dados de origem e destino na fase de treinamento, adicionando um recurso binário que desempenha o papel de sinalizador indicando o domínio da amostra de dados. Essa abordagem é provavelmente a estratégia de DA (Detecção de Anomalias) supervisionada mais simples concebível, em que as probabilidades previstas de fraude de ambos os classificadores são calculadas usando uma média aritmética ponderada, em que o peso α (com $\alpha \in [0, 1]$) é ajustado por validação **(LEBICHOT, 2021).**

Já a Análise de Componentes Principais (PCA) tem como objetivo reduzir a dimensionalidade de conjuntos de dados multivariados. Em outras palavras, busca-se representar de maneira reduzida um conjunto de variáveis aleatórias, sem que isso acarrete perda significativa de informações. Matematicamente, as Componentes Principais (PC) são combinações lineares não correlacionadas das variáveis aleatórias **(HONGYU, SANDANIELO e JUNIOR, 2015).**

Parte dos dados utilizados no trabalho de **LEBICHOT (2021)**, foram disponibilizados em base pública no site kaggle.com.

O conjunto de dados contém transações feitas por cartões de crédito em setembro de 2013 por titulares de cartões europeus. Esses dados são transações ocorridas em dois dias, com 492 registros de fraudes em 284.807 transações. O conjunto de dados é altamente desbalanceado, a classe positiva (fraudes) responde por 0,172% de todas as transações.

Todas as variáveis de entrada são numéricas resultado de uma transformação PCA. Não foram fornecidas mais informações sobre os dados originais, nem o significado das variáveis transformadas. As variáveis são nomeadas de V1, V2, … V28 que são os principais componentes obtidos com PCA, possui ainda as variáveis 'Time' e 'Amount'.

O recurso 'Time' contém os segundos decorridos entre cada transação e a primeira transação no conjunto de dados. O recurso 'Amount' é o valor da transação realizada. A característica 'Class' é a variável resposta e assume valor 1 em caso de fraude e 0 quando não fraude **KAGGLE (2010).**

A base de dados possui 31 variáveis sendo:

- Vinte oito das variáveis sendo o resultado de transformação PCA e nomeadas com a letra V, numerada de 1 a 28, ou seja, V1, V2, V3, ..., V28.
- Uma variável registrando o valor das transações realizadas, identificada como Amount.
- Uma variável Class com registro 0 e 1, onde, 0 representa registros de não fraude e 1 representa registros de ocorrência de fraude.
- Uma variável registrando a diferença entre a transação corrente registrada por cada linha da base de dado se o momento da primeira transação da base dados. Assim, a variável Time representa o tempo entre a transação atual e a primeira transação da base.

A base de dados possui um total de 284.807 registros, sendo 284.315 registros de transações não fraudulentas (99,8272%) e 492 registros de transações fraudulentas (0,1728).

Um ponto importante a ser considerado no cenário de fraude em cartões de crédito é que o evento estudado é extremamente raro em relação ao volume total de transações realizadas. Dessa forma, as bases para identificação desses eventos apresentam a mesma característica.

7. METODOLOGIA

Nesta seção serão apresentadas as descrições das metodologias utilizadas para desenvolvimento desse trabalho. Serão abordados os conceitos teóricos desses métodos, bem como as fórmulas e medidas de ajuste utilizadas.

7.1. Modelo Logístico

7.1.1. Aplicação do Modelo Logístico

Foi utilizado o modelo logístico que é uma técnica estatística utilizada para modelar a relação entre uma variável binária dependente e um conjunto de variáveis independentes. Neste trabalho, o objetivo é detectar fraudes, ou seja, identificar se uma observação pertence à classe "fraude" ou "não fraude".

O modelo logístico é muito utilizado em casos onde há necessidade de realização de Análise Discriminante que é uma técnica estatística multivariada utilizada para classificar observações em grupos pré-definidos com base em um conjunto de variáveis preditoras. Seu objetivo é encontrar uma função discriminante que maximize a separação entre os grupos e minimize a variabilidade dentro de cada grupo.

A função logística é a função de ligação utilizada no modelo logístico. Ela mapeia a combinação linear das variáveis independentes para um valor entre 0 e 1, representando a probabilidade de ocorrência do evento de interesse (fraude, neste caso). A forma geral do modelo logístico é dada por:

$$P(Y = 1) = 1 - \frac{1}{(1 + \exp(\beta 0 + \beta 1X1 + \beta 2X2 + \cdots + \beta pXp)}$$

ou seja,

$$P(Y = 1) = \frac{\exp(\beta 0 + \beta 1X1 + \beta 2X2 + \cdots + \beta pXp}{(1 + \exp(\beta 0 + \beta 1X1 + \beta 2X2 + \cdots + \beta pXp)}$$

onde:

- $P(Y = 1)$ – é a probabilidade condicional de ocorrência da classe “fraude” dado um conjunto de variáveis independentes X. A variável Y pode assumir os valores 0 e 1 sendo 0 não fraude e 1 fraude.
- $\beta 0, \beta 1, \beta 2 \ldots \beta p$ – são os coeficientes do modelo, que representam a contribuição de cada variável independente na probabilidade de ocorrência de fraude.
- $X1, X2, \ldots, Xp$ – são as variáveis independentes.

7.1.2. A Função de verossimilhança

O modelo logístico utiliza a função de verossimilhança na otimização de seus parâmetros. A função de verossimilhança na regressão logística é uma medida da probabilidade dos dados observados. A fórmula matemática para a função de verossimilhança é a seguinte:

$$L(\beta) = \prod_{i=1}^{n} pi^{yi} * (1 - pi)^{1-yi}$$

onde:

- $L(\beta)$ - representa a função de verossimilhança.
- β - é um vetor de parâmetros a serem estimados.
- $pi = 1 - \frac{1}{(1+\exp(\beta 0+\beta 1X1i+\beta 2X2i+\cdots+\beta pXpi)}$ - é a probabilidade condicional de ocorrência da classe “fraude” dado um conjunto de variáveis independentes X observadas para cada elemento amostral i, i=1,2,...,n.
- yi - pode assumir os valores 0 e 1 sendo 0 não fraude e 1 fraude. É o valor observado para cada elemento amostral 1, i=1,2,...,n.
- $\prod_{i=1}^{n}$ - representa o produto de todas as observações.

A função de verossimilhança L(β) é definida como o produto das funções de densidade de probabilidade (ou massa de probabilidade) para cada observação, assumindo que as observações são independentes. O objetivo é encontrar os valores dos parâmetros que maximizem essa função, ou seja, que tornem os dados observados mais prováveis de acordo com o modelo. Na prática, para evitar problemas numéricos com produtos de muitos valores pequenos, é comum trabalhar com a função de log-verossimilhança em vez da função de verossimilhança.

7.1.3. Métricas de avaliação

O ajuste do modelo logístico gera algumas medidas que permitem avaliar o ajuste do modelo e a significância das variáveis independentes utilizadas. Essas medidas são descritas a seguir:

a) **Null deviance**:

É a medida de discrepância do modelo sem a inclusão de nenhum preditor**.** Em um contexto de regressão logística, o modelo nulo atribui a mesma probabilidade de ocorrência para todas as observações, independentemente dos valores dos preditores. A fórmula é descrita por:

$$Null\ deviance = -2 * \log(L_0)$$

onde:

- A *Null deviance* é calculada como -2 vezes o logaritmo da função de verossimilhança para o modelo nulo, que é um modelo sem preditores, ou seja, um modelo que não inclui nenhuma variável independente além do intercepto. Portanto, a única estimativa de parâmetro no modelo nulo é o intercepto, representado por β_0. Ao calcular a null deviance, estamos avaliando a adequação desse modelo simplificado em descrever os dados, considerando apenas a variabilidade explicada pelo intercepto.
- $\log(L_0)$ – é o logaritmo da função de verossimilhança para o modelo nulo, que é um modelo sem preditores.

b) **Residual deviance**:

É uma medida da discrepância entre o modelo ajustado e os dados observados. Também é usado para avaliar o ajuste do modelo. Quanto menor a residual deviance, melhor o ajuste do modelo

$$Residual\ deviance = \mp\left\{2\sum_{i=1}^{n}[y_i\ \log(p_i) + (1 -\ y_i)\log(1 -\ p_i)]\right\}^{1/2}$$

onde:

- n – é o número total de observações no conjunto de dados.
- y_i - é o valor observado da variável resposta (0 ou 1, onde 0 significa "não fraude" e 1 significa "fraude").
- $pi = 1 - \frac{1}{(1+\exp(\beta 0+\beta 1X1i+\beta 2X2i+\cdots+\beta pXpi)}$ que é a probabilidade condicional de ocorrência da classe "fraude" dada a observação i, descrito na fórmula da seção 7.1.1 deste trabalho.

"Hosmer, Lemeshow e Sturdivant (2013**)** discutem a residual deviance como uma métrica de ajuste de modelos logísticos, enfatizando seu papel na avaliação da

qualidade do ajuste em relação aos dados observados. A fórmula da deviance é apresentada como uma medida baseada na log-verossimilhança que facilita comparações entre modelos, aproximando-se de uma distribuição qui-quadrado para modelos bem ajustados" **(Hosmer, Lemeshow & Sturdivant, 2013, pp. 143-146)**.

Essa fórmula é uma medida de verossimilhança negativa escalada e, ao ser multiplicada por 2, torna a residual deviance comparável à distribuição qui-quadrado, o que é útil em testes de ajuste e para comparação entre modelos.

Importante ressaltar que pi não pode ser exatamente igual a zero (nem a 1) em um modelo de regressão logística padrão. A função logística mapeia qualquer valor real para uma probabilidade entre 0 e 1, mas nunca alcança exatamente 0 ou 1.

c) **AIC (Akaike Information Criterion)**:

É um critério de seleção para comparação de diferentes modelos estatísticos. Ele fornece uma medida que leva em consideração o melhor ajuste do modelo e a menor complexidade, levando em consideração a quantidade de informação capturada pelos dados e o número de parâmetros estimados **(Emiliano, 2009).**

$$AIC = -2 * \log(L) + 2 * k$$

onde:

- $\log(L)$ – valor numérico do logaritmo da função de verossimilhança obtido para o modelo ajustado.
- k - é o número de parâmetros estimados no modelo.

O objetivo do AIC é encontrar o modelo que melhor equilibra a capacidade de ajustar os dados ($-2 * \log(L)$ e a complexidade do modelo ($2 * k$). O modelo com o menor valor de AIC é considerado o melhor em termos de balancear esses dois aspectos.

d) **Matriz Confusão**:

Além das medidas citadas são utilizadas outras que permitem avaliar a capacidade para de acerto do modelo tanto no grupo de interesse (fraude) quanto no grupo (não fraude) como por exemplo a matriz de confusão. Ela é composta por quatro elementos:

- Verdadeiro Positivo (VP): número de observações corretamente classificadas como "fraude".
- Verdadeiro Negativo (VN): número de observações corretamente classificadas como "não fraude".
- Falso Positivo (FP): número de observações incorretamente classificadas como "fraude" (falsos alarmes).
- Falso Negativo (FN): número de observações incorretamente classificadas como "não fraude" (fraudes não detectadas).

A Tabela 1 apresenta a disposição dos elementos na matriz confusão:

Tabela 1 – Representação matriz confusão

		Real	
		Não Fraude	Fraude
Predito	Não Fraude	VN	FP
	Fraude	FN	VP

Fonte: Elaboração Própria

Na Tabela 1 tem-se as colunas que representam os rótulos reais dos dados (fraude e sem fraude), e as linhas representam as previsões do modelo (fraude e sem fraude).

Com base na matriz de confusão, podem ser calculados os percentuais de acerto (ou taxa de acerto) e erro do modelo, como descrito a seguir:

e) **Acurácia do modelo:**

Ela permite avaliar a precisão geral do modelo. Embora seja uma métrica importante, ela pode ser enganosa em problemas de desequilíbrio de classes pois o modelo pode ser altamente preciso ao prever transações legítimas e ainda assim não ser eficaz na detecção de fraudes.

$$ACR = (VP + VN)/(VP + VN + FP + FN)$$

onde:

- ACR–a acurácia permite medir proporção de previsões corretas em relação ao total de previsão.

f) **Precisão do modelo:**

Esta medida permite avaliar qual a capacidade do modelo de explicar o evento de fraude. No contexto de prevenção a fraudes uma alta precisão significa que o modelo está sendo cuidadoso ao rotular transações como fraude, evitando classificar incorretamente transações legítimas como fraude.

$$PRC = VP/(VP + FP)$$

onde:

- PRC – a precisão permite medir proporção de previsões corretas de fraude em relação a quantidade total de casos de fraude previstos.

g) **Sensibilidade do modelo:**

A sensibilidade avalia a precisão do modelo, ou seja, a capacidade de prever a fraude sem impactar transações sem fraude. A sensibilidade é crucial em modelos de prevenção a fraude, pois indica a capacidade do modelo em detectar a fraude real. Um alto valor de sensibilidade indica que o modelo está sendo capaz de identificar a maioria das transações fraudulentas, minimizando os falsos negativos.

$$SEN = VP/(VP + FN)$$

onde:

- SEN - a sensibilidade permite medir proporção de previsões corretas de fraude em relação a quantidade total de casos de fraude reais.

h) **Especificidade do modelo:**

A especificidade é relevante em modelos de prevenção a fraude, pois indica a capacidade do modelo em evitar rotular erroneamente transações legítimas como fraude. Um alto valor de especificidade significa que o modelo está sendo preciso na identificação de transações legítimas, minimizando os falsos positivos.

$$ESP = VN/(VP + FP)$$

onde:

- ESP - a especificidade permite medir proporção de previsões corretas de sem fraude em relação a quantidade total de casos de sem fraudes reais.

7.2. Análise de Componentes Principais

7.2.1. Aplicação da Análise de Componentes Principais

A Análise de Componentes Principais (PCA) é uma técnica estatística utilizada para simplificar a complexidade de conjuntos de dados multidimensionais, identificando as principais direções ao longo das quais os dados variam. Ela é frequentemente utilizada em análise exploratória de dados, redução de dimensionalidade e visualização de dados, além disso, permite agrupar os indivíduos com base na variação de suas características dentro da população, representando assim o comportamento variável dos indivíduos em relação ao conjunto de características que os define **(Varella,2008).**

De acordo com **Mingoti (2013)**, “PCA tem como objetivo criar combinações lineares de um conjunto de variáveis originais com base na estrutura de variância e covariância de um vetor aleatório, composto por p variáveis aleatórias”.

O resultado dessas combinações, chamado de componentes principais, gera um conjunto de novas variáveis criadas a partir das variáveis originais, geralmente em menor quantidade do que o grupo original.

Para compreensão dos resultados obtidos a partir da realização da estimação PCA é necessário a definição de alguns conceitos:

- Variância total: é a soma da variância das p-variáveis de estudo. No modelo PCA busca-se encontrar o conjunto de componentes principais que consigam explicar o maior percentual da variância total utilizando-se o menor número de componentes para isto.
- Autovalores: são as variâncias dos componentes principais e representam a quantidade de variância dos dados que cada componente principal explica.
- Percentual de variância explicada: é o percentual da variância total dos dados que cada componente principal explica.
- Variância total explicada: é o resultado da soma da variância de cada componente principal, após a escolha das componentes relevantes que serão utilizadas, dividido pela Variância total.
- As correlações das variáveis com as componentes principais: esta medida representa a relação entre as variáveis originais e as componentes principais e auxiliam na interpretação do significado das componentes principais.
- Os escores das componentes: são os valores resultantes da transformação dos dados originais em uma nova base das componentes principais.

Nas etapas que se seguem buscou-se descrever de forma sucinta as etapas necessárias para a construção do PCA.

7.2.2. Etapas da aplicação da Análise de Componentes Principais

a) **Organização da Matriz dos dados**:

Antes de iniciar o PCA, os dados devem ser organizados em uma matriz onde cada linha representa uma observação (indivíduo ou objeto) e cada coluna representa uma variável. Assim a matriz de dados X terá dimensão n x p, onde n é o número de observações e p é o número de variáveis **(Varella,2008).**

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ x_{31} & x_{32} & \cdots & x_{3p} \\ x_{41} & x_{42} & \cdots & x_{4p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

Na Matriz X observa-se a representação da matriz de dados organizada.

b) **Padronização dos dados**:

Um aspecto crucial no processo de construção da Análise de Componentes Principais (PCA) é a consideração da escala das variáveis. Embora a padronização seja comumente empregada para garantir a comparabilidade entre elas, vale ressaltar que nem sempre é imprescindível. A padronização consiste em subtrair a média de cada variável e dividir pelo desvio padrão, contudo, a sensibilidade da PCA à escala das variáveis deve ser avaliada caso a caso, levando em conta a natureza dos dados e os objetivos da análise **(Varella, 2008)**.

$$\boldsymbol{Z_{ij} = (x_{ij} - \overline{x}_j)/s_j}$$

onde:

- i - representa o índice da observação amostra
- j - representa o índice da variável
- Z_{ij}- é o valor da j-ésima variável do i-ésimo elemento amostral após a padronização.
- x_{ij} - É o valor observado da j-ésima variável do i-ésimo elemento amostral.
- $\overline{x}_j$ – É a média amostral dos valores da j-ésima variável.
- s_j – É o desvio padrão amostral dos valores da j-ésima variável.

c) **Cálculo da matriz de covariâncias:**

A matriz de covariâncias é uma matriz quadrada que contém as covariâncias entre pares de variáveis em um conjunto de dados multivariados. A covariância é uma medida importante que descreve como as variáveis estão relacionadas umas com as outras em termos de dispersão conjunta, isto é, variância e covariância **(Varella, 2008)**.

A covariância entre duas variáveis X_l e X_j é uma medida de como essas duas variáveis variam juntas. Uma covariância positiva indica que, quando uma variável aumenta de valor, a outra também tende a aumentar, enquanto uma covariância

negativa indica que quando uma variável aumenta de valor, a outra tende a diminuir **Varella (2008)**.

A matriz de covariâncias é sempre uma matriz simétrica de dimensão p x p, onde p é o número de variáveis analisadas. A diagonal principal da matriz representa a variância individual de cada variável enquanto os demais valores representam a covariância entre as variáveis **Varella (2008).**

A fórmula para calcular a covariância entre duas variáveis $\mathbf{X_l}$ e $\mathbf{X_j}$ é dada respectivamente por:

$$S_{lj} = \frac{\sum_{i=1}^{n}(X_{il} - \bar{X}_l)\left(X_{ij} - \bar{X}_j\right)}{n-1}$$

Onde:

- S_{lj}- Covariância entre as variáveis x e z.
- n - Número total de observações na amostra.
- X_{il} - Valor da i-ésima observação da variável X_l.
- X_{ij} - Valor da i-ésima observação da variável X_j.
- $\bar{X}_l$ - Média amostral da variável X_l.
- $\bar{X}_j$ - Média amostral da variável X_j.

A variância é uma medida estatística que quantifica a dispersão de uma variável em relação à sua média. Para a j-ésima variável, a variância é definida como a média dos quadrados das diferenças entre cada observação e a média da variável.

A variância fornece informações sobre a distribuição dos dados em torno da média. Uma variância alta indica que os dados estão dispersos em relação à média, enquanto uma variância baixa sugere que os dados estão mais concentrados em torno da média.

A matriz de covariâncias das variáveis originais (denotada usualmente por $\sum_{pxp}$) será estimada pela matriz de covariâncias amostral S_{pxp} definida por:

$$S_{Pxp} = \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1p} \\ S_{21} & S_{22} & \cdots & S_{2p} \\ S_{31} & S_{32} & \cdots & S_{3p} \\ S_{41} & S_{42} & \cdots & S_{4p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{P1} & S_{p2} & \cdots & S_{pp} \end{bmatrix}$$

d) **Cálculo da matriz de correlação**:

A matriz de correlação é calculada a partir da matriz de covariâncias, mas com a diferença fundamental de que todos os valores são normalizados para estar na faixa de -1 a 1, o que facilita a interpretação e a comparação entre variáveis. Valores de correlação positivos mais próximos de 1 indicam uma alta correlação positiva entre as variáveis, enquanto que valores de correlações negativos mais próximos de -1 alta correlação negativa entre as variáveis. Os valores próximos a 0 indicam baixa correlação entre as variáveis **(Varella,2008).**

Cada elemento da matriz representa o coeficiente de correlação entre pares de variáveis. A diagonal principal é sempre preenchida com 1, pois a correlação de uma variável consigo mesma é sempre perfeita **(Varella,2008).**

A matriz de correlação é obtida normalizando a matriz de covariâncias. Para calcular o elemento (l, j) da matriz de correlação, a covariância entre as variáveis $\mathrm{X_l}$ e $\mathrm{X_j}$ é dividida pela raiz quadrada do produto das variâncias individuais de l e j dessas variáveis.

A fórmula para calcular a correlação entre duas variáveis X_l e X_j é:

$$\mathrm{R_{lj}} = \frac{\mathrm{S_{lj}}}{\sqrt{\mathrm{S_{ll}S_{jj}}}}$$

Esse é o coeficiente de correlação amostral entre as l-ésima e j-ésima variáveis.

A matriz de correlação de estimada R_{pxp} será denotada por:

$$R_{Pxp} = \begin{bmatrix} R_{11} & R_{12} & \cdots & R_{1p} \\ R_{21} & R_{22} & \cdots & R_{2p} \\ R_{31} & R_{32} & \cdots & R_{3p} \\ R_{41} & R_{42} & \cdots & R_{4p} \\ \vdots & \vdots & \ddots & \vdots \\ R_{P1} & R_{p2} & \cdots & R_{pp} \end{bmatrix}$$

e) **Cálculo dos autovalores e autovetores:**

A partir da definição das matrizes de Covariância ou Correlação é possível determinar os autovalores. Os autovalores geram um coeficiente para cada variável da matriz de covariância e esses coeficientes são utilizados na construção da combinação linear do correspondente ao autovalor. Os autovetores representam as direções principais do conjunto de dados, e os autovalores indicam a variabilidade explicada por cada componente principal. Os autovetores correspondentes aos maiores autovalores representam as direções de maior variação nos dados originais. Essas direções são usadas para reduzir a dimensionalidade dos dados, mantendo a maior parte da informação **(Silva, 2008).**

Existem muitos exemplos na literatura que utilizam a técnica PCA, nos quais o leitor poderá encontrar exemplos da aplicação e dos cálculos matemáticos, como em **Mingoti (2013)** capítulos 2 e 3.

8. ANÁLISE DESCRITIVA

8.1. Análise exploratória dos dados

Na análise descritiva dos dados utilizou-se toda a base de dados conforme descrito na seção 6. Foi realizada a validação do preenchimento dos dados e verificou-se que não havia dados faltantes. Nas Tabelas de 1 - 5 temos as estatísticas descritivas de cada variável explicativa, com intuito se avaliar as principais medidas de dispersão e amplitude para essas variáveis.

Tabela 2– Análise Descritiva variáveis Time, V1, V2, V3, V4 e V5

Class	Medidas	Time	V1	V2	V3	V4	V5
SEM FRAUDE	Min.	0	-56,41	-72,72	-48,33	-5,68	-113,74
	1st Qu.	54230	-0,92	-0,60	-0,88	-0,85	-0,69
	Median	84711	0,02	0,06	0,18	-0,02	-0,05
	Mean	94838	0,01	-0,01	0,01	-0,01	0,01
	3rd Qu.	139333	1,32	0,80	1,03	0,74	0,61
	Max.	172792	2,45	18,90	9,38	16,88	34,80
	Desvio Padrão	5	1,93	1,64	1,46	1,40	1,36
	Assimetria	0	-3,13	-4,89	-1,45	0,58	-2,21
	Curtose	-1	31,06	98,40	14,54	2,10	217,85
Class	Medidas	Time	V1	V2	V3	V4	V5
FRAUDE	Min.	406	-30,55	-8,40	-31,10	-1,31	-22,11
	1st Qu.	41242	-6,04	1,19	-8,64	2,37	-4,79
	Median	75569	-2,34	2,72	-5,08	4,18	-1,52
	Mean	80747	-4,77	3,62	-7,03	4,54	-3,15
	3rd Qu.	128483	-0,42	4,97	-2,28	6,35	0,21
	Max.	170348	2,13	22,06	2,25	12,12	11,10
	Desvio Padrão	5	6,78	4,29	7,11	2,87	5,37
	Assimetria	0	-1,79	1,22	-1,51	0,49	-1,35
	Curtose	-1	2,81	2,54	1,73	-0,22	1,68

Fonte: Elaboração Própria

Dentre os principais pontos verificados na análise descritiva da Tabela 2 destaca-se:

- A transação não fraude tem um tempo médio maior que o tempo de fraude, variável Time.
- As variáveis V1, V3 e V5 possuem valores dos quartis, da média e da mediana menores nos registros de fraude em relação a não fraude.
- A variável V2 e V4 possuem valores quartis, da média e da mediana maior nos registros de fraude em relação a não fraude.
- Nos dados de assimetria destaca-se a variável V2 que apresenta valor negativo para registros não fraude e positivo para registros de fraude.
- Na análise de curtose destaca-se as variáveis V1, V2, V3 e V5 com valores distintos entre fraude e não fraude.

Tabela 3– Análise Descritiva variáveis V6, V7, V8, V9, V10 e V11

Class	Medidas	V6	V7	V8	V9	V10	V11
SEM FRAUDE	Min.	-26,16	-31,76	-73,22	-6,29073	-14,74	-4,8
	1st Qu.	-0,77	-0,55	-0,21	-0,64	-0,53	-0,76
	Median	-0,27	0,04	0,02	-0,05	-0,09	-0,03
	Mean	0,00	0,01	0,00	0,00	0,01	-0,01
	3rd Qu.	0,40	0,57	0,33	0,60	0,46	0,74
	Max.	73,30	120,59	18,71	15,59	23,75	10,00
	Desvio Padrão	1,33	1,18	1,16	1,09	1,04	1,00
	Assimetria	1,84	4,75	-8,40	0,67	2,42	0,15
	Curtose	42,99	452,91	209,77	3,20	21,96	0,04
Class	Medidas	V6	V7	V8	V9	V10	V11
FRAUDE	Min.	-6,41	-43,56	-41,04	-13,43	-24,59	-1,70
	1st Qu.	-2,50	-7,97	-0,20	-3,87	-7,76	1,97
	Median	-1,42	-3,03	0,62	-2,21	-4,58	3,59
	Mean	-1,40	-5,57	0,57	-2,58	-5,68	3,80
	3rd Qu.	-0,41	-0,95	1,76	-0,79	-2,61	5,31
	Max.	6,47	5,80	20,01	3,35	4,03	12,02
	Desvio Padrão	1,86	7,21	6,80	2,50	4,90	2,68
	Assimetria	0,86	-1,82	-2,81	-0,97	-1,15	0,50
	Curtose	2,81	4,16	16,50	1,45	1,41	0,18

Fonte: Elaboração Própria

Dentre os principais pontos verificados na análise descritiva da Tabela 3 destaca-se:

- As variáveis V6, V7, V9 e V10 possuem valores dos quartis, da média e da mediana menores nos registros de fraude em relação a não fraude.
- A variável V8 possui média e mediana maiores nos registros de fraude em relação a registros de não fraude.
- A variável V11 possuem valores dos quartis, da média e da mediana maior nos registros de fraude em relação a não fraude.
- Nos dados de assimetria destaca-se a variável V7 que apresenta valor negativo para registros fraude e positivo para registros de não fraude.
- Na análise de curtose destaca-se as variáveis V6, V7, V8 e V10 com valores distintos entre fraude e não fraude.

Tabela 4– Análise Descritiva variáveis V12, V13, V14, V15, V16 e V17

Class	Medidas	V12	V13	V14	V15	V16	V17
SEM FRAUDE	Min.	-15,14	-5,79	-18,39	-4,39	-10,12	-17,10
	1st Qu.	-0,40	-0,65	-0,42	-0,58	-0,47	-0,48
	Median	0,14	-0,01	0,05	0,05	0,07	-0,06
	Mean	0,01	0,00	0,01	0,00	0,01	0,01
	3rd Qu.	0,62	0,66	0,49	0,65	0,52	0,40
	Max.	7,85	7,13	10,53	8,88	17,32	9,25
	Desvio Padrão	9,46	9,95	8,97	9,15	8,45	7,49
	Assimetria	-1,26	0,07	-0,70	-0,31	-0,44	0,22
	Curtose	5,30	0,20	7,72	0,28	2,85	20,05
Class	Medidas	V12	V13	V14	V15	V16	V17
FRAUDE	Min.	-18,68	-3,13	-19,21	-4,50	-14,13	-25,16
	1st Qu.	-8,69	-0,98	-9,69	-0,64	-6,56	-11,95
	Median	-5,50	-0,07	-6,73	-0,06	-3,55	-5,30
	Mean	-6,26	-0,11	-6,97	-0,09	-4,14	-6,67
	3rd Qu.	-2,97	0,67	-4,28	0,61	-1,23	-1,34
	Max.	1,38	2,82	3,44	2,47	3,14	6,74
	Desvio Padrão	4,65	1,10	4,28	1,05	3,87	6,97
	Assimetria	-0,66	-0,03	-0,25	-0,53	-0,49	-0,48
	Curtose	-0,22	-0,45	-0,30	0,63	-0,47	-0,46

Fonte: Elaboração Própria

Dentre os principais pontos verificados na análise descritiva da Tabela 4 destaca-se:

- As variáveis V12, V14, V16 e V17 possuem valores dos quartis, da média e da mediana menores nos registros de fraude em relação a não fraude. Estas variáveis possuem também um valor de desvio-padrão maior para registros de fraude em relação a não fraude.
- As variáveis V13 e V15 possuem o primeiro quartil, a média e a mediana menores nos registros de fraude em relação a não fraude. Já o terceiro quartil possui valores semelhantes para fraude e não fraude.
- Os dados de assimetria apresentam pequenas oscilações entre fraude e não fraude sem grandes destaques para uma variável específica.
- Na análise de curtose destaca-se as variáveis V12, V14 e V17 com valores distintos entre fraude e não fraude.

Tabela 5– Análise Descritiva variáveis V18, V19, V20, V21, V22 e V23

Class	Medidas	V18	V19	V20	V21	V22	V23
SEM FRAUDE	Min.	-5,37	-7,21	-54,50	-34,83	-10,93	-44,81
	1st Qu.	-0,50	-0,46	-0,21	-0,23	-0,54	-0,16
	Median	0,00	0,00	-0,06	-0,03	0,01	-0,01
	Mean	0,00	0,00	0,00	0,00	0,00	0,00
	3rd Qu.	0,50	0,46	0,13	0,19	0,53	0,15
	Max.	5,04	5,59	39,42	22,61	10,50	22,53
	Desvio Padrão	8,25	8,12	7,69	7,17	7,24	6,22
	Assimetria	-0,04	0,10	-2,08	3,01	-0,19	-5,81
	Curtose	0,84	1,70	273,29	176,81	2,25	444,23
Class	Medidas	V18	V19	V20	V21	V22	V23
FRAUDE	Min.	-9,50	-3,68	-4,13	-22,80	-8,89	-19,25
	1st Qu.	-4,66	-0,30	-0,17	0,04	-0,53	-0,34
	Median	-1,66	0,65	0,28	0,59	0,05	-0,07
	Mean	-2,25	0,68	0,37	0,71	0,01	-0,04
	3rd Qu.	0,09	1,65	0,82	1,24	0,62	0,31
	Max.	3,79	5,23	11,06	27,20	8,36	5,47
	Desvio Padrão	2,90	1,54	1,35	3,87	1,49	1,58
	Assimetria	-0,51	0,04	2,15	2,64	-1,49	-5,36
	Curtose	-0,53	-0,21	16,46	30,61	17,90	65,85

Fonte: Elaboração Própria

Dentre os principais pontos verificados na análise descritiva da Tabela 5 destaca-se:

- A variável V18 possui valor do 1º quartil, da média e da mediana menores nos registros de fraude em relação a não fraude e possui desvio padrão maior em fraude em relação a não fraude.
- As variáveis V19, V20, V22, V23 não possui variável que se destaque entre fraude e não fraude.
- A variável V21 possui valor do quartil, da média e da mediana maior nos registros de fraude em relação a não fraude e possui desvio padrão maior em fraude em relação a não fraude.
- Nos dados de assimetria destaca-se a variável V20 que apresenta valor negativo para registros não fraude e positivo para registros de fraude.
- Na análise de curtose destaca-se as variáveis V20, V21, V22 e V23 com valores distintos entre fraude e não fraude.

Tabela 6– Análise Descritiva variáveis V24, V25, V26, V27, V28 e Amount

Class	Medidas	V24	V25	V26	V27	V28	Amount
SEM FRAUDE	Min.	-2,84	-10,30	-2,60	-22,57	-15,43	-
	1st Qu.	-0,35	-0,32	-0,33	-0,07	-0,05	5,65
	Median	0,04	0,02	-0,05	0,00	0,01	22,00
	Mean	0,00	0,00	0,00	0,00	0,00	88,29
	3rd Qu.	0,44	0,35	0,24	0,09	0,08	77,05
	Max.	4,58	7,52	3,52	31,61	33,85	25.691,16
	Desvio Padrão	6,06	5,21	4,82	4,00	3,30	2,50
	Assimetria	-0,55	-0,41	0,58	-1,07	11,27	17,00
	Curtose	0,62	4,27	0,92	252,53	940,82	846,72
Class	Medidas	V24	V25	V26	V27	V28	Amount
FRAUDE	Min.	-2,03	-4,78	-1,15	-7,26	-1,87	-
	1st Qu.	-0,44	-0,31	-0,26	-0,02	-0,11	1,00
	Median	-0,06	0,09	0,00	0,39	0,15	9,25
	Mean	-0,11	0,04	0,05	0,17	0,08	122,21
	3rd Qu.	0,29	0,46	0,40	0,83	0,38	105,89
	Max.	1,09	2,21	2,75	3,05	1,78	2.125,87
	Desvio Padrão	5,16	7,97	4,72	1,38	5,47	2,57
	Assimetria	-0,45	-0,78	0,53	-2,25	-0,71	3,73
	Curtose	0,06	3,63	1,93	7,98	1,45	17,47

Fonte: Elaboração Própria

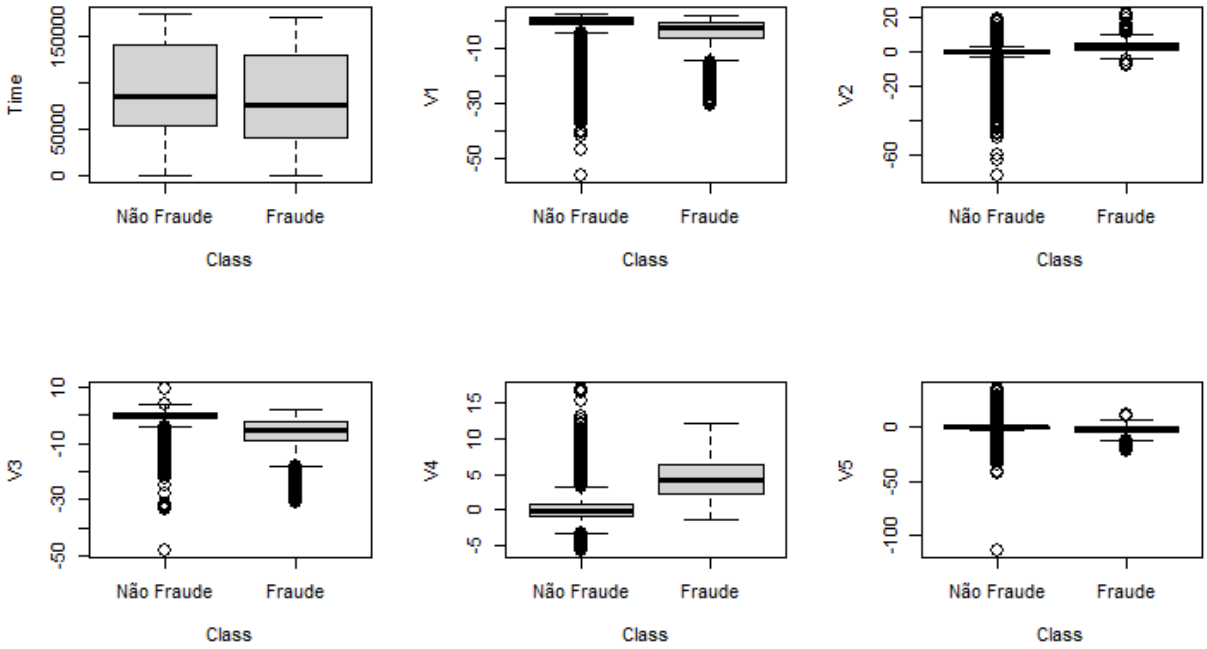
Dentre os principais pontos verificados na análise descritiva da Tabela 6 destaca-se:

- Nos dados de assimetria destaca-se a variável V28 que apresenta valor negativo para registros fraude e positivo para registros de não fraude.
- Na análise de curtose destaca-se as variáveis V27, V28 e Amount com valores distintos entre fraude e não fraude.
- Não identificamos outros destaques para estas variáveis.

8.2. Análise gráfica dos dados

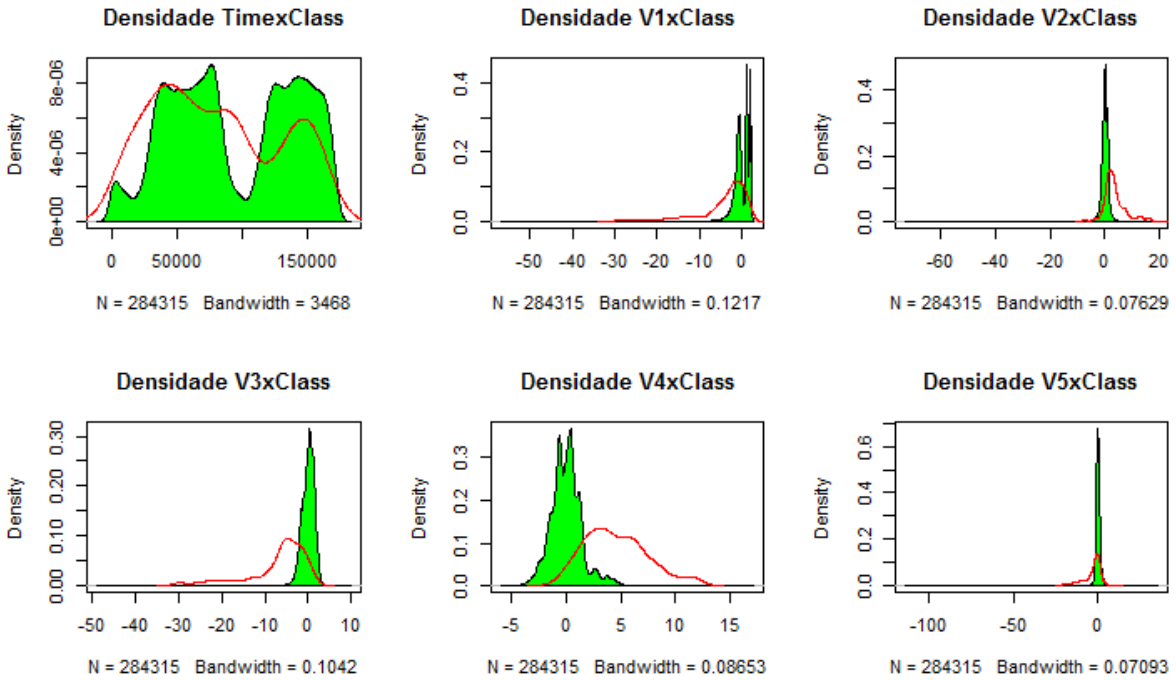
Nos Quadros 1-10 que se seguem, tem-se uma análise boxplot e a análise de densidade que buscam avaliar a capacidade de separação entre registros de fraude e não fraude de forma individual de cada variável.

Quadro 1 - Boxplot para variáveis Time, V1, V2, V3, V4 e V5



Fonte: Elaboração Própria

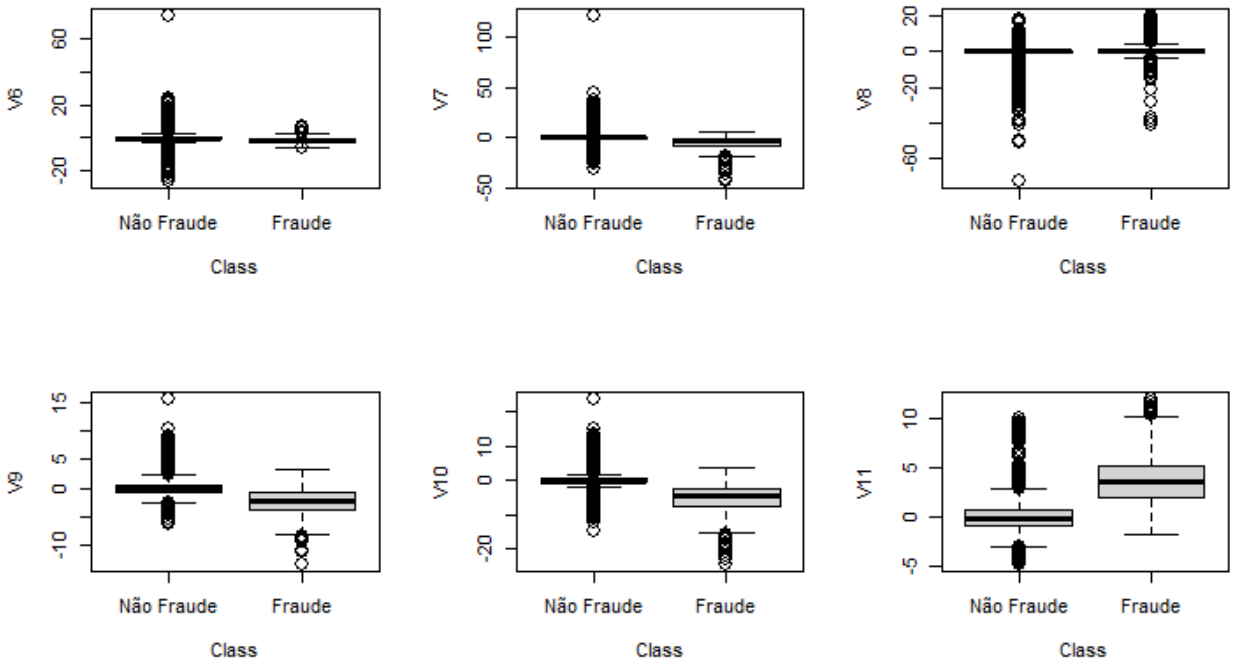
Quadro 2 - Densidade Fraude x Não Fraude para variáveis Time, V1, V2, V3, V4 e V5



Fonte: Elaboração Própria

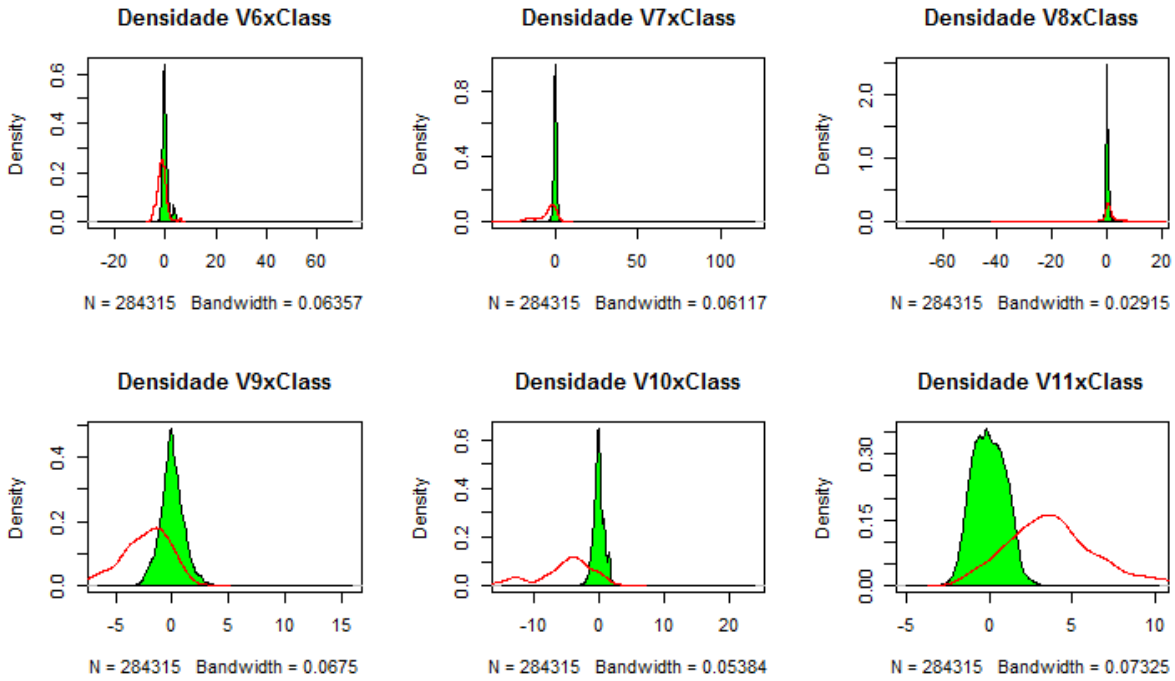
De acordo com os Quadros 1 e 2 as variáveis V3 (média fraude: -7,033 | não fraude: 0,01217) e V4 (média fraude: 4,542 | não fraude: -0,00786) apresentaram médias mais distintas entre sie a análise de densidade demonstra que essas possuem maior separação com a variável resposta. A variável V2 apresenta uma menor dispersão dos dados em relação a variável resposta com baixa separação.

Quadro 3 – Boxplot para variáveis V6, V7, V8, V9, V10 e V11



Fonte: Elaboração Própria

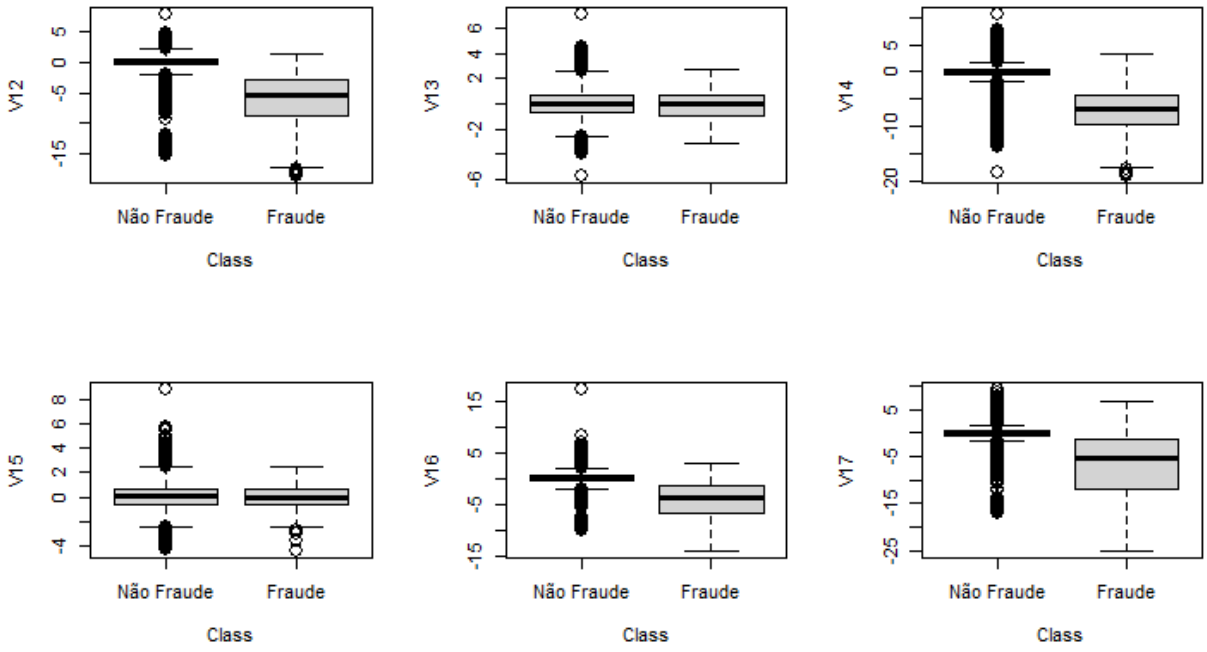
Quadro 4- Densidade Fraude x Não Fraude para variáveis V6, V7, V8, V9, V10 e V11



Fonte: Elaboração Própria

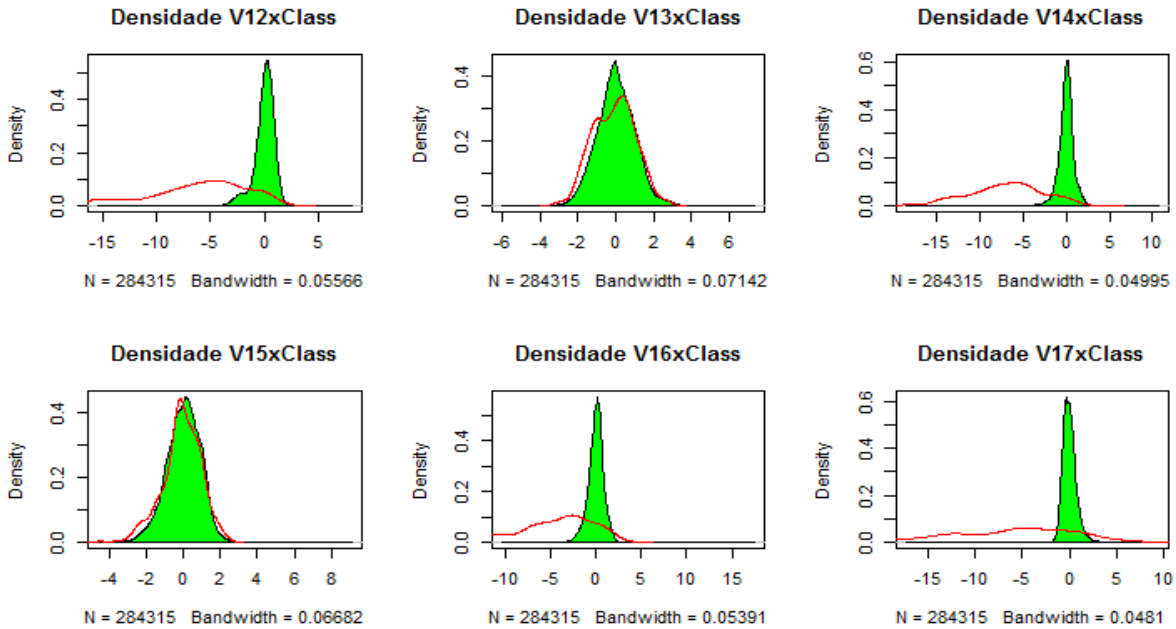
De acordo com os Quadros 3 e 4 as variáveis V9 (média fraude: -2,5811 | não fraude:0,004467), V10 (média fraude: -5,677 | não fraude: 4,542) e V11 (média fraude: 3,800 | não fraude: --0,006576) apresentaram maior separação entre variável explicativa e variável resposta.

Quadro 5- Boxplot para variáveis V12, V13, V14, V15, V16 e V17



Fonte: Elaboração Própria

Quadro 6- Densidade Fraude x Não Fraude para variáveis V12, V13, V14, V15, V16 e V17

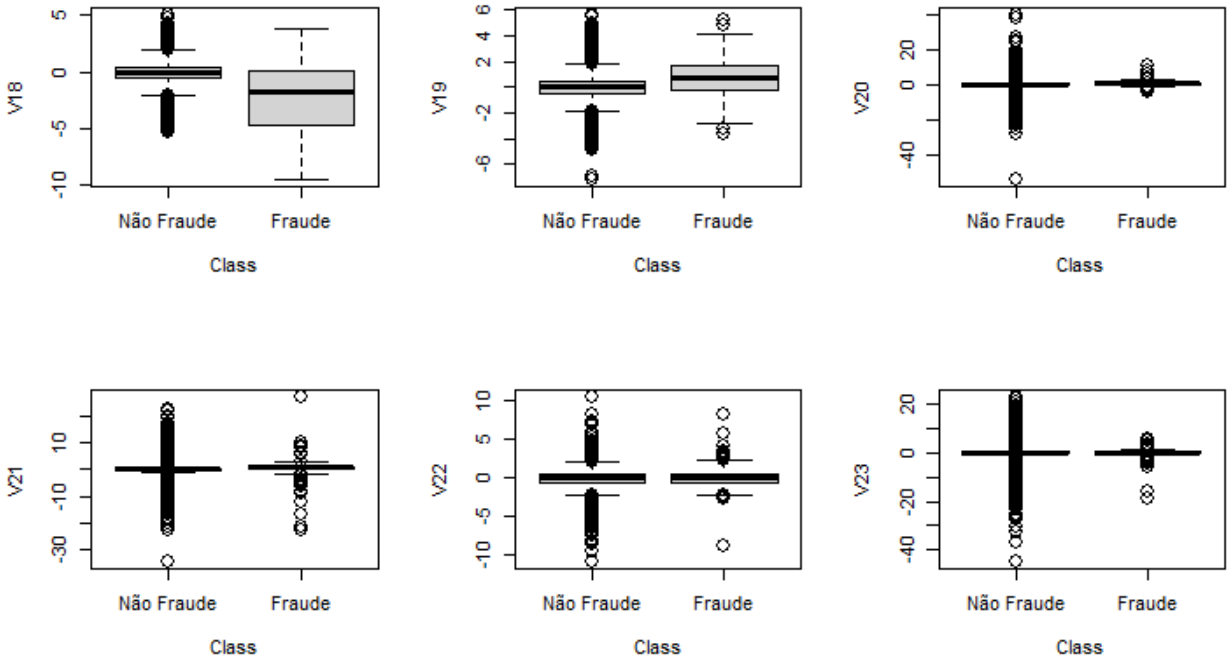


Fonte: Elaboração Própria

De acordo com os Quadros 5 e 6 as variáveis V12 (média fraude: -6,259 | não fraude: 0,01083), V14 (média fraude: -6,972 | não fraude: 0,01206) e V16 (média fraude: -4,140 | não fraude: 0,007164) apresentaram maior relevância, com maior

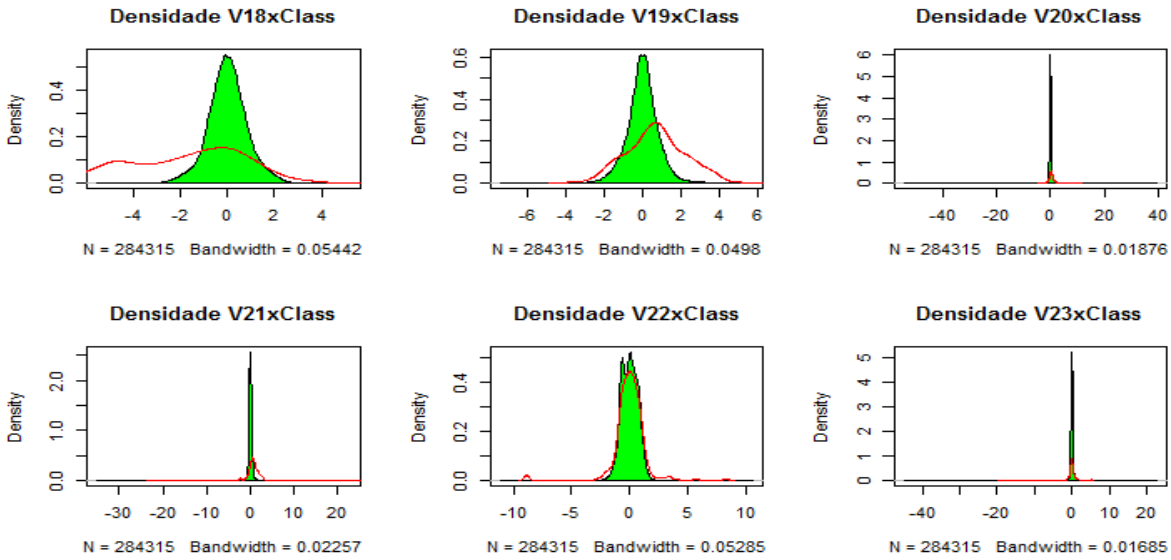
separação entre fraude e não fraude, além de apresentar médias mais distintas. A variável V17 (média fraude: -6,666 | não fraude: 0,01154).

Quadro 7- Boxplot para variáveis V18, V19, V20, V21, V22 e V23



Fonte: Elaboração Própria

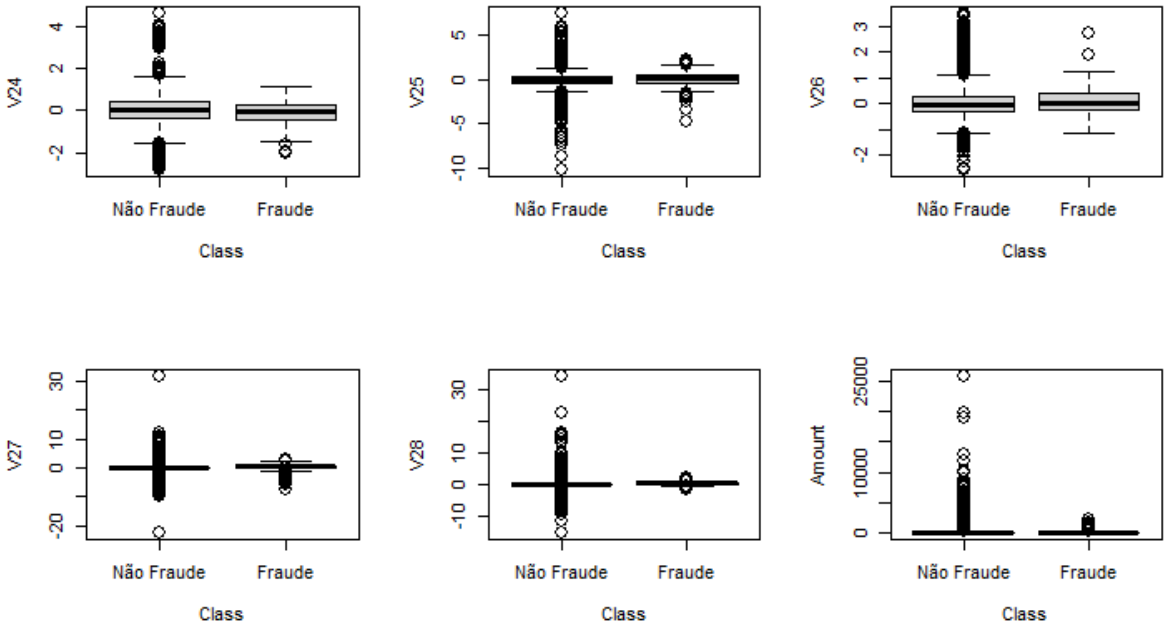
Quadro 8- Densidade Fraude x Não Fraude para variáveis V18, V19, V20, V21, V22 e V23



Fonte: Elaboração Própria

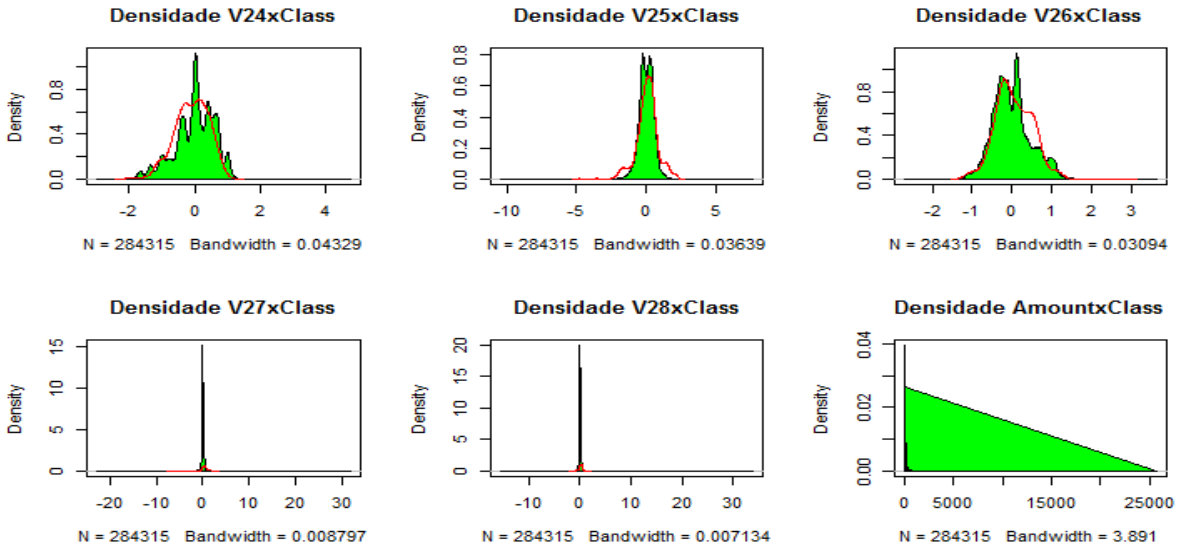
Nos Quadros7 e 8 apenas V18 e V19 apresentaram separação entre fraude e não fraude. A variável V18 (média fraude: fraude -2,24631 | não fraude: 0,003887) e uma maior dispersão para os dados de fraude (desvio padrão: fraude 2,8993| não fraude: 8,2491). A variável V19 (média fraude: fraude 0,6807 | não fraude: -0,001178) apresentou médias mais distantes entre fraude e não fraude.

Quadro 9- Boxplot para variáveis V24, V25, V26, V27, V28 e Amount



Fonte: Elaboração Própria

Quadro 10- Densidade Fraude x Não Fraude para variáveis V24, V25, V26, V27, V28 e Amount



Fonte: Elaboração Própria

De acordo com os Quadros 9 a 10 apenas a variável Amount tem concentração de média igual para explicativa e resposta. Contudo, a dispersão dos dados para não fraude é menor com uma média relativamente maior (média fraude: fraude 122,21 | não fraude: 88,29).

9. MODELO PREDITIVO

9.1. Análise de correlação das variáveis

A análise de correlação é fundamental no processo de análise de dados, pois permite identificar relações lineares entre as variáveis explicativas que podem comprometer os resultados das estimativas do modelo logístico alterando os resultados finais. A matriz de correlação contém medida estatística que indica como duas variáveis estão relacionadas entre si e a intensidade desta relação.

A Tabela 7 apresenta os resultados da matriz de correlação.

Tabela 7 - Matriz de correlação entre variáveis explicativas

	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20	V21	V22	V23	V24	V25	V26	V27	V28	AMO
Time	100%																													
V1	24%	100%																												
V2	-22%	-82%	100%																											
V3	14%	88%	-87%	100%																										
V4	-22%	-62%	68%	-77%	100%																									
V5	29%	86%	-79%	84%	-58%	100%																								
V6	7%	32%	-28%	45%	-41%	25%	100%																							
V7	21%	88%	-84%	89%	-71%	81%	30%	100%																						
V8	-13%	-8%	-2%	-17%	10%	-20%	-56%	8%	100%																					
V9	15%	66%	-70%	76%	-79%	65%	37%	76%	-8%	100%																				
V10	21%	74%	-77%	86%	-80%	75%	41%	86%	-5%	85%	100%																			
V11	-32%	-54%	62%	-71%	80%	-52%	-48%	-64%	17%	-69%	-80%	100%																		
V12	28%	60%	-67%	76%	-84%	62%	49%	72%	-16%	76%	88%	-90%	100%																	
V13	-14%	-4%	1%	-5%	4%	-12%	-8%	-1%	26%	-1%	-4%	5%	-8%	100%																
V14	17%	45%	-57%	66%	-80%	43%	52%	54%	-18%	68%	75%	-89%	87%	0%	100%															
V15	-17%	12%	-16%	14%	-12%	9%	-5%	18%	14%	13%	16%	-5%	7%	-3%	0%	100%														
V16	23%	64%	-64%	73%	-73%	69%	42%	75%	-17%	73%	85%	-81%	90%	-10%	78%	0%	100%													
V17	24%	69%	-65%	74%	-72%	74%	42%	77%	-22%	76%	85%	-77%	88%	-12%	73%	4%	95%	100%												
V18	26%	68%	-62%	70%	-64%	74%	35%	76%	-18%	71%	80%	-67%	80%	-11%	62%	3%	90%	94%	100%											
V19	-7%	-31%	22%	-33%	31%	-41%	-21%	-36%	22%	-34%	-42%	41%	-46%	15%	-38%	25%	-63%	-59%	-56%	100%										
V20	-5%	-33%	31%	-36%	30%	-29%	-14%	-40%	-3%	-36%	-37%	20%	-23%	-5%	-13%	-13%	-19%	-22%	-21%	6%	100%									
V21	-6%	1%	4%	3%	-2%	5%	2%	4%	-12%	16%	9%	14%	-7%	0%	-22%	14%	-15%	-9%	-8%	12%	-52%	100%								
V22	15%	-3%	-3%	-4%	9%	-8%	1%	-10%	3%	-23%	-19%	-1%	-8%	2%	9%	-10%	-8%	-11%	-10%	12%	42%	-75%	100%							
V23	6%	-5%	16%	-4%	2%	-9%	31%	-10%	-44%	-5%	-5%	-2%	1%	-10%	1%	-5%	-1%	1%	1%	-1%	13%	13%	0%	100%						
V24	-2%	-6%	0%	1%	-8%	-15%	-1%	-3%	7%	4%	0%	-13%	4%	4%	15%	1%	-5%	-8%	-11%	11%	-3%	-6%	8%	-2%	100%					
V25	-19%	-7%	14%	-9%	-2%	-10%	-13%	5%	22%	0%	2%	2%	2%	0%	-9%	1%	4%	3%	7%	-15%	2%	12%	-25%	5%	-6%	100%				
V26	-2%	6%	0%	-3%	16%	6%	-4%	2%	5%	-12%	-4%	17%	-12%	3%	-19%	4%	-6%	-5%	-2%	6%	4%	3%	1%	2%	-11%	8%	100%			
V27	-13%	19%	-16%	11%	-3%	18%	-14%	25%	29%	14%	17%	17%	-2%	4%	-20%	18%	-2%	1%	6%	5%	-14%	36%	-36%	-19%	-19%	20%	19%	100%		
V28	3%	18%	1%	14%	-9%	19%	-6%	15%	-1%	16%	15%	2%	1%	-12%	-12%	9%	2%	6%	9%	-6%	3%	30%	-25%	7%	-4%	17%	5%	28%	100%	
AMOU	-1%	-2%	-20%	0%	1%	-16%	26%	16%	1%	3%	1%	-2%	2%	4%	4%	5%	-1%	-1%	-1%	8%	5%	1%	0%	-15%	6%	-10%	-2%	11%	-13%	100%

Fonte: Elaboração Própria

Verifica-se existência de correlação positiva entre as variáveis Time, V1, V3, V4, V5, V6, V7, V9, V10, V11, V12, V13, V14, V16, V17, V18, V19, V20, V23, V25,

V27, V28 e Amount que apresentaram pelo menos uma correlação positiva superiora 0,20 com pelo menos uma variável.

Este fato representa um problema para construção do modelo logístico pois, a presença de multicolinearidade produz instabilidade nas estimativas desse modelo gerando resultados inexatos, conforme dito por **Nakamura (2013).**

> A multicolinearidade ocorre quando existe uma relação exata ou aproximada entre as colunas da matriz de planejamento do modelo, fazendo com que seja introduzida uma singularidade aproximada nessa matriz e, consequentemente, produzindo instabilidade nas estimativas dos parâmetros obtidas através do método de máxima verossimilhança, nos seus respectivos erros padrão e nos resultados de alguns testes estatísticos. Como consequência desse fato, as conclusões e a inferência sobre os parâmetros baseada neste modelo poderão ficar seriamente comprometidas. Seu correto diagnóstico e técnicas para amenizar e até eliminar o problema e as suas consequências em modelos de regressão logística são raras na literatura **(NAKAMURA, 2013).**

A presença de multicolinearidade representa um problema no ajuste dos modelos de regressão logística, pois esse, têm como pré-requisito que as variáveis explicativas não possuam relação entre si.

Para solucionar o problema de multicolinearidade (**MOHANTY; MAHAPATRA, 2018**) indica que:

> Para lidar com a multicolinearidade, existem várias abordagens possíveis. Uma delas é remover variáveis altamente correlacionadas do modelo. Outra opção é combinar as variáveis correlacionadas em uma única variável, por exemplo, calculando uma média ou soma ponderada das variáveis. A análise de componentes principais (PCA) é outra técnica útil para lidar com multicolinearidade. A PCA transforma as variáveis originais em um novo conjunto de variáveis não correlacionadas, chamadas de componentes principais. Por fim, é possível utilizar técnicas de regularização, como a regressão Ridge ou Lasso, que adicionam uma penalização aos coeficientes das variáveis no modelo, reduzindo a influência de variáveis correlacionadas **(MOHANTY e MAHAPATRA, 2018).**

Diante das soluções propostas por **MOHANTY e MAHAPATRA (2018)** opta-se pela realização de duas soluções que são:

- A remoção de variáveis altamente correlacionadas para o ajuste do modelo logístico.
- A aplicação de componentes principais PCA com a criação de um novo grupo de variáveis.

Nesse ponto, é importante ressaltar que a técnica PCA foi aplicada no estudo original de **Lebichot (2021)**, conforme descrito na seção 6. Contudo, os dados e as

variáveis do estudo original são mais volumosos e os critérios de avaliação utilizados levaram em consideração as restrições e o objetivo proposto para aquele estudo. Desta forma é importante destacar a diferença entre a aplicação da PCA no estudo original versus o trabalho atual, pois, o presente trabalho utiliza parte da base já transformada e um menor volume de dados o que muda **a** importância das componentes em termos de explicação de variância total, sendo válido a utilização dessa técnica para ajuste dos dados.

Ao final, o modelo de regressão logística será desenvolvido para cada uma das soluções aqui apresentadas e os resultados comparados.

9.2. Base de dados variáveis originais

Como forma de determinar o grau de relevância das variáveis explicativas em relação a variável resposta utilizou-se de o pacote *glmnet versão 4.1.1* do *Software R versão 4.2.3*. O pacote *glmnet versão 4.1.1* **(Friedman, Hastie, e Tibshirani, 2020)** é uma implementação em R de algoritmos de regressão linear com regularização Lasso e Ridge. A regularização Lasso pode ser usada para seleção de variáveis, uma vez que tende a forçar os coeficientes das variáveis menos importantes a zero. Já a regularização Ridge pode ser usada para lidar com problemas de multicolinearidade, uma vez que ela reduz os coeficientes das variáveis altamente correlacionadas para valores próximos de zero **(Farias,2021).**

O pacote *glmnet* implementa essas técnicas em um contexto de regressão linear, mas também pode ser usado em modelos de regressão logística. Além disso, o pacote fornece ferramentas para selecionar o valor ideal do parâmetro de regularização usando validação cruzada **(Friedman, Hastie, e Tibshirani, 2020).**

O resultado do "coeficiente estimado para cada variável" é uma lista que contém um vetor de valores estimados para cada valor de lambda. O parâmetro lambda é um fator de ajuste que controla o nível de penalização aplicado aos coeficientes das variáveis preditoras no modelo. Quanto maior o valor de lambda, maior será a penalização e mais próximos de zero serão os coeficientes do modelo. Cada vetor de coeficientes representa a importância de cada variável na regressão logística para um determinado valor de lambda. O lambda selecionado é aquele que minimiza o erro médio quadrático (MSE) na validação cruzada.

A aplicação do pacote *glmnet* versão 4.1.1 resulta em coeficiente que é interpretado como a contribuição da variável explicativa na previsão da variável resposta (binária). Um coeficiente positivo indica que a variável tem uma relação positiva com a variável resposta, enquanto um coeficiente negativo indica uma relação negativa. Um coeficiente próximo de zero indica que uma variável não contribui significativamente para a previsão da variável resposta **(Friedman, Hastie, e Tibshirani, 2010**).

A Tabela 8 foi criada com a união das variáveis explicativas, o maior valor de correlação apresentado por estas variáveis e o coeficiente estimado para cada variável, através do *glmnet*. Foi criada uma ordenação das variáveis explicativas considerando o grau de correlação e o resultado da aplicação do pacote *glmnet* versão 4.1.1 (Coeficiente Estimado) para retirada das variáveis que apresentam maior correlação com outra variável e menor relevância.

Tabela 8– Ordem da retirada das variáveis correlacionadas

Variavel	Maior Correlação	Coeficiente Estimado	Status
Time	0,29	-1,21	Permanece
V1	0,88		Removido
V2	0,68		Removido
V3	0,89		Removido
V4	0,80	7,67	Permanece
V5	0,81	1,25	Removido
V6	0,52	-1,48	Permanece
V7	0,86		Removido
V8	0,29	-1,94	Permanece
V9	0,85	-8,94	Permanece
V10	0,88	-4,12	Removido
V11	0,70	1,94	Removido
V12	0,90	-5,00	Removido
V13	0,17	-3,01	Permanece
V14	0,87	-6,16	Removido
V15	0,25	-6,95	Permanece
V16	0,95	-9,82	Removido
V17	0,94		Removido
V18	0,86		Removido
V19	0,34		Permanece
V20	0,42	-1,82	Permanece
V21	0,36	1,31	Permanece
V22	0,27	1,37	Permanece
V23	0,22	-3,39	Permanece
V24	0,12		Permanece
V25	0,20	9,69	Permanece
V26	0,19	-3,98	Permanece
V27	0,31		Permanece
V28	0,25	5,11	Permanece
AMOUNT	0,25	9,77	Permanece

Fonte: Elaboração Própria

Após análise da Tabela 8 as variáveis eliminadas foram:

- Variáveis V1, V2, V3, V5, V7, V10, V11, V12, V14, V16, V17 e V18 foram retiradas devido a multicolinearidade.
- Variáveis Time, V4, V6, V8, V9, V13, V15, V19, V20, V21, V22, V23, V24, V25, V26, V27, V28 e Amount apresentaram baixa correlação permanecendo.

Na Tabela 9 apresenta-se o resultado da matriz de correlação das variáveis que permaneceram.

Tabela 9 - Matriz de correlação após ajuste de multicolinearidade

	Time	V4	V6	V8	V9	V13	V15	V19	V20	V21	V22	V23	V24	V25	V26	V27	V28	Amount
Time	1.00																	
V4	-0.16	1.00																
V6	0.05	-0.34	1.00															
V8	-0.08	0.09	-0.41	1.00														
V9	0.10	-0.67	0.29	-0.08	1.00													
V13	-0.08	0.02	-0.04	0.17	-0.03	1.00												
V15	-0.17	-0.08	0.00	0.09	0.09	-0.03	1.00											
V19	-0.05	0.30	-0.17	0.17	-0.28	0.10	0.08	1.00										
V20	-0.05	0.25	-0.09	-0.02	-0.30	-0.01	-0.05	0.08	1.00									
V21	-0.02	0.02	0.00	-0.09	0.10	0.00	0.07	0.11	-0.39	1.00								
V22	0.12	0.08	0.00	0.03	-0.16	-0.01	-0.03	0.07	0.27	-0.58	1.00							
V23	0.06	0.01	0.22	-0.37	-0.03	-0.07	-0.05	-0.01	0.07	0.07	0.00	1.00						
V24	-0.02	-0.06	-0.01	0.04	0.01	0.02	-0.02	0.06	-0.01	-0.04	0.03	-0.01	1.00					
V25	-0.22	-0.02	-0.08	0.17	-0.02	-0.01	-0.02	-0.09	0.00	0.08	-0.13	0.00	-0.01	1.00				
V26	-0.01	0.11	-0.01	0.03	-0.06	0.02	0.00	0.05	0.01	0.02	-0.01	0.01	-0.01	0.00	1.00			
V27	-0.08	0.01	-0.11	0.22	0.08	0.02	0.08	0.04	-0.10	0.31	-0.27	-0.17	-0.11	0.14	0.09	1.00		
V28	-0.01	-0.06	-0.08	0.03	0.08	-0.05	0.02	-0.01	0.11	0.25	-0.20	0.05	-0.03	0.09	-0.01	0.13	1.00	
Amount	-0.03	0.03	0.20	-0.01	-0.02	-0.01	0.02	-0.01	0.25	0.05	-0.06	-0.17	0.01	-0.08	0.00	0.06	-0.06	1.00

Fonte: Elaboração Própria

9.3. Definição da Amostragem dos dados

Conforme registrado na seção 6 o evento de fraude em cartões de crédito é um evento extremamente raro dentro das transações totais, o que faz com que o desafio de identificação seja ainda maior.

A base inicial dos dados apresenta 284.807 registros sendo 492 registros de transações fraudulentas o que representa 0,1728%.

O balanceamento da base é algo importante, pois, como a maioria das observações pertence à classe Sem Fraude, o modelo poderá tender a prever a classe majoritária em todas as situações gerando uma alta acurácia, mas não será útil na prática, pois não está capturando as nuances das outras classes Fraude.

Em bases de dados desbalanceadas é comum que técnicas tradicionais para determinar o tamanho da amostra não sejam adequadas. Assim definiu-se a utilização da técnica de amostragem estratificada, para realização da amostragem dos dados. Esta técnica divide a população em estratos de acordo com a distribuição das classes e, em seguida, uma amostra é selecionada de cada estrato. **(YAN, YE e CHEN (2016).**

Optou-se pela redução do desbalanceamento da base original, mantendo-se uma proporção aproximada de quatro registros sem fraude para cada um registro de fraude, desta forma, evita-se o risco de o modelo aprender apenas as características da classe majoritária, mas ao mesmo tempo, mantém-se as características associadas a natureza rara do evento.

Importante ressaltar que seria possível a utilização de uma base balanceada para execução desde trabalho, desde que respeitadas as características observadas na população, contudo, a utilização de uma base que mantém um grau de desbalanceamento permite que a base de dados mantenha uma distribuição proporcionalmente semelhante a população.

Utilizou-se para composição da base os 492 registros de fraude existentes na base original e, utilizando a função sample 3.6.2 nativa versão 6.0-86 *do Software R versão 4.2.3* **(BECKER, CHAMBERS e WILKS (1988)),** foi realizada uma estratificação nos 284.315 registros sem fraude para composição dos registros sem fraude.

A Tabela 10 demonstra a distribuição final da amostra.

Tabela 10 – Resumo da Base Amostral

Base amostral		
Class.	Qtd.	%
0	1889	79,3%
1	492	20,7%
Total	2381	100%

Fonte: Elaboração Própria

9.4. Modelo 1 – regressão logística para variáveis originais

Na primeira etapa será desenvolvido um modelo de regressão logística sobre a base de dados já previamente tratada conforme descrito na seção 9.3 que apresenta as variáveis identificadas como relevantes para o ajuste do modelo.

A base de dados possui 19 variáveis descritas como Time, V4, V6, V8, V9, V13, V15, V19, V20, V21, V22, V23, V24, V25, V26, V27, V28, Amount e Class, sendo "Class" a variável resposta e as demais variáveis explicativas.

Inicialmente, esta base foi separada em dois grupos: um grupo para realização do treino do modelo e outro grupo para realização dos testes. Esse procedimento é realizado para que o modelo seja treinado em uma base aprendendo as características necessárias, e posteriormente realizada uma validação na base de teste para confirmação dos resultados apresentados no treino.

Conforme descrito na seção 9.3 e na Tabela 10, a base inicial possui 2381 registros, sendo 492 registros de fraude (20,7%) e 1889 registros de não fraude (79,3%).

Para realização da estratificação da base foi utilizado o pacote *caret* versão 6.0.93 *do Software R versão 4.2.3*. que utiliza uma função que retorna um vetor de índices que indicam quais observações devem ser utilizadas para treinamento e quais devem ser utilizadas para teste **(Kuhn,2010).**

A base de treino teve (70%) dos registros, e a base de teste teve (30%). Elas apresentam as seguintes distribuições:

Base Treino: fraude - 332 (20%) | não fraude –1335(80%) registros;

Base Teste: fraude - 160 (22%) | não fraude - 554 (78%) registros.

Foi ajustado o modelo logístico com todas as variáveis selecionadas (ver Tabela 10), obtendo-se os resultados apresentados na Quadro 11.

Quadro 11- Estimativas dos parâmetros do modelologístico variáveis originais com todas as variáveis da Tabela 10.

```
Call:
glm(formula = Class ~ ., family = binomial, data = treino)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3292  -0.2494  -0.1223  -0.0191   3.3829

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.803e+00  3.247e-01 -11.709  < 2e-16 ***
Time        -8.516e-06  2.853e-06  -2.985 0.002831 **
V4           1.087e+00  8.786e-02  12.373  < 2e-16 ***
V6          -1.137e+00  1.328e-01  -8.562  < 2e-16 ***
V8          -3.769e-01  6.751e-02  -5.583 2.36e-08 ***
V9          -5.375e-01  1.408e-01  -3.818 0.000135 ***
V13         -4.067e-01  1.199e-01  -3.393 0.000692 ***
V15         -1.790e-01  1.382e-01  -1.295 0.195262
V19          3.182e-01  1.236e-01   2.575 0.010030 *
V20         -3.837e-01  1.527e-01  -2.513 0.011980 *
V21          4.404e-01  9.817e-02   4.487 7.24e-06 ***
V22          4.322e-01  1.847e-01   2.341 0.019257 *
V23          1.725e-01  1.144e-01   1.507 0.131740
V24         -9.481e-01  2.481e-01  -3.822 0.000132 ***
V25          3.346e-01  2.266e-01   1.476 0.139854
V26         -6.724e-01  2.816e-01  -2.388 0.016961 *
V27          8.662e-02  2.208e-01   0.392 0.694869
V28          7.320e-01  2.679e-01   2.733 0.006280 **
Amount       1.624e-03  4.289e-04   3.787 0.000152 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1664.45  on 1666  degrees of freedom
Residual deviance:  522.64  on 1648  degrees of freedom
AIC: 560.64

Number of Fisher Scoring iterations: 7
```

Fonte: Elaboração Própria

“.” – Considerar ponto como separador de decimal.

A coluna Pr (>|z) do Quadro 11 mostra o p-value (probabilidade de significância, p- valor) para cada coeficiente estimado. Esse valor é associado ao

teste de hipótese para cada coeficiente no modelo. Se o p-value (p-valor) for menor do que um determinado nível de significância (geralmente 0,05), a hipótese nula é rejeitada, e o coeficiente é considerado estatisticamente significativo.

As variáveis Time, V4, V6, V8, V9, V13, V19, V20, V21, V22, V24, V26, V28 e Amount, destacadas no Quadro 11, apresentaram resultados estatisticamente significativos para o modelo pois tiveram o p-value menor que 0,05. Já as variáveis V15, V23, V25 e V27 apresentaram p-value muito superiores a 0,05, não se rejeitando assim a hipótese nula, ou seja, não são estatisticamente significativas.

O valor do erro padrão (Std. Error) auxilia na avaliação das variáveis independentes quanto à precisão da estimativa do parâmetro, pois representa a variabilidade da estimativa do coeficiente.

Tem-se ainda como medidas de qualidade do modelo a Null deviance: 1664,45, que representa a variância total do modelo a ser explicada quando nenhuma variável explicativa é incluída. A Residual deviance: 522,64 representa a variância residual a ser explicada após a inclusão das variáveis explicativas no modelo. Assim, a diferença entre estas duas medidas indica a capacidade de explicação do modelo atual. Quanto menor o Residual deviance, maior é a capacidade total de explicação do modelo.

Outra medida importante de avaliação do modelo é o AIC (Akaike Information Criterion), que é uma medida de avaliação de modelos que leva em consideração a qualidade de ajuste e a complexidade do modelo. Esta medida permite escolher o modelo que melhor explica e é menos complexo, pois penaliza modelos com mais variáveis. Assim, um modelo que possua um AIC significativamente menor do que o outro modelo é considerado o melhor modelo. O AIC inicial do modelo é 560,64

Por fim, tem-se o Number of Fisher Scoring iterations (7), que é o número de iterações realizadas durante o processo de estimação dos parâmetros do modelo estatístico. Ele é usado para maximização da função de verossimilhança. No caso, o número seis representa o número de iterações para se obter a solução de máxima verossimilhança **(BRESLOW, 1987)**.

Assim, serão utilizados os valores do p-value, erro padrão, Null deviance, Residual deviance e AIC para comparação entre o resultado dos modelos citados na seção 7.

Após a remoção das variáveis V15, V23, V25 e V27 e o novo ajuste do modelo a variável V22 apresentou resultado estatisticamente não significativo com p-value de 0,0932, sendo assim removida do modelo (saída computacional não apresentada).

No Quadro 12 apresenta-se os resultados do ajuste retirando-se as variáveis V15, V22, V23, V25 e V27.

Quadro 12- Estimativas dos parâmetros do modelo logístico variáveis originais otimizado.

```
Call:
glm(formula = Class ~ V4 + V6 + V8 + V9 + V13 + V19 + V20 + V21
    V24 + V26 + V28 + Amount, family = binomial, data = treino)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2468  -0.2706  -0.1473  -0.0244   3.5473

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.3483670  0.2331333 -18.652  < 2e-16 ***
Time        -8.516e-06  2.853e-06  -2.985 0.002831 **
V4           1.0365213  0.0834371  12.423  < 2e-16 ***
V6          -1.0380212  0.1275178  -8.140 3.95e-16 ***
V8          -0.3593867  0.0664848  -5.406 6.46e-08 ***
V9          -0.5890128  0.1361626  -4.326 1.52e-05 ***
V13         -0.4062713  0.1167998  -3.478 0.000505 ***
V19          0.3142178  0.1142753   2.750 0.005966 **
V20         -0.3140372  0.1551366  -2.024 0.042943 *
V21          0.3521083  0.0895134   3.934 8.37e-05 ***
V24         -0.7804431  0.2314690  -3.372 0.000747 ***
V26         -0.6703862  0.2650932  -2.529 0.011443 *
V28          0.6607372  0.3139340   2.105 0.035317 *
Amount       0.0012403  0.0004036   3.073 0.002118 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1664.45  on 1666  degrees of freedom
Residual deviance:  542.98  on 1654  degrees of freedom
AIC: 568.98

Number of Fisher Scoring iterations: 7
```

Fonte: Elaboração Própria
“.” – Considerar ponto como separador de decimal.

Após o ajuste do modelo com a retirada das variáveis não estatisticamente significativas verifica-se que não ocorreram alterações significativas nos resultados gerais do modelo (Residual deviance = 542,98). Também se vê que o AIC teve pouca variação (AIC= 568,98).

Mantidas as variáveis significativas a 5% para o modelo, passa-se a predição da base de teste com apuração dos resultados.

9.5. Resultados modelo de regressão logística para variáveis originais

No Quadro 13 são apresentados os resultados obtidos no treino e no teste do modelo destacando abaixo:

- A Taxa de detecção de fraude no momento do Treino foi de 76,80% se confirmando na base de Teste com 70,62%. (Sensibilidade)
- A Especificidade no momento do Treino foi de 97,82% se confirmando na base de Teste com 99,27%.
- A Taxa do Falso Alarme no momento do Treino foi de 2,17% passando para 0,72% na base de teste.

Quadro 13- Matriz confusão e resultados do modelo variáveis originais treino e teste

Matriz de Confusão -Base Treino

		Real		
		Não Fraude	Fraude	Tot.
Predito	Não Fraude	1306	77	1383
	Fraude	29	255	284
	Tot	1335	332	1667

Proporções de Acertos

Não Fraude	Fraude
0,97828	0,76807

Proporções de Erros

Não Fraude	Fraude
0,02172	0,23193

Proporções de Acertos totais

0,93641

Proporções de Erros totais

0,06359

Matriz de Confusão -Base Teste

		Real		
		Não Fraude	Fraude	Tot.
Predito	Não Fraude	550	47	597
	Fraude	4	113	117
	Tot	554	160	714

Proporções de Acertos

Não Fraude	Fraude
0,99278	0,70625

Proporções de Erros

Não Fraude	Fraude
0,00722	0,29375

Proporções de Acertos totais

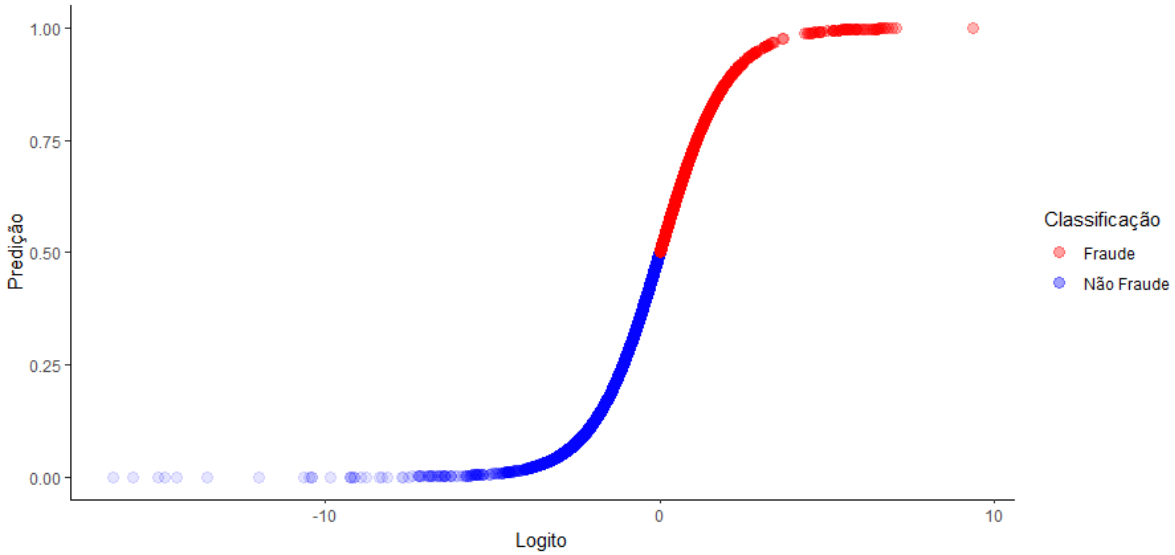
0,92857

Proporções de Erros totais

0,07143

Fonte: Elaboração Própria

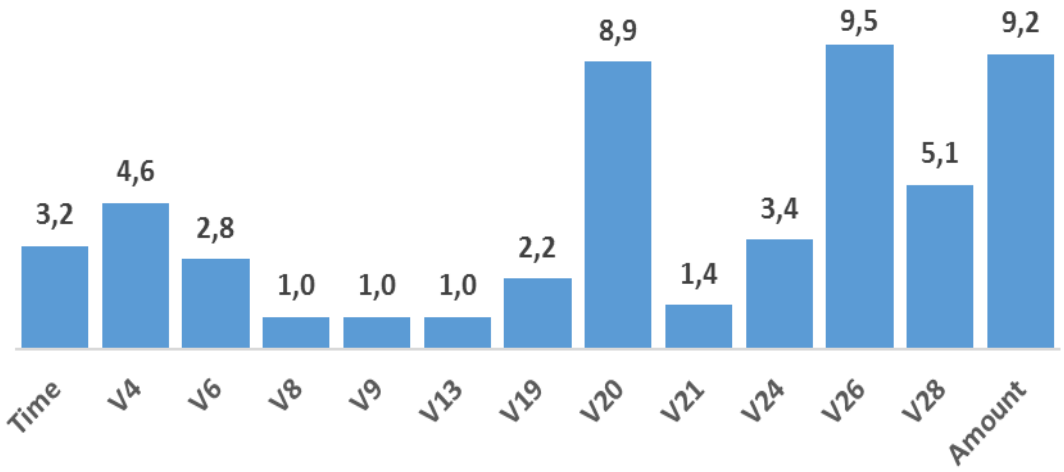
Gráfico 5 - Gráfico da curva logística variáveis originais



Fonte: Elaboração Própria

O Gráfico 5 demonstra o bom ajuste do modelo logístico separando os registros de fraude e não fraude e atribuindo uma maior probabilidade de risco para os casos de fraude.

Gráfico 6 - Gráfico da Razão de Chances variáveis originais



Fonte: Elaboração Própria

A razão de chances é calculada como a razão entre as chances de o evento ocorrer em um grupo, ou seja, é a probabilidade de o evento acontecer dividida pela probabilidade de o evento não acontecer.

A interpretação da razão de chances é baseada no valor resultante. Se o valor for igual a 1, indica que não há diferença nas chances de o evento ocorrer entre os dois grupos. Se for maior que 1, indica que as chances de o evento ocorrer são maiores no grupo exposto em relação ao grupo de referência.

O Gráfico 6 demonstra a razão de chances para cada variável do modelo permitindo determinar a contribuição de cada variável a partir de sua ocorrência.

É importante ressaltar que a razão de chances não estabelece uma relação causal entre as variáveis, apenas descreve a associação entre elas. Para determinar relações causais, são necessários estudos experimentais ou outros métodos de análise.

A equação do modelo ajustado de regressão logística é dada pela função logística, que relaciona a probabilidade de ocorrência de um evento (variável dependente) às variáveis explicativas (variáveis independentes). Para o modelo ajustado no Quadro 12 esta equação é descrita como:

$$P(Y = 1) = -4{,}3483670 - 0{,}0000008516 * Time + 1{,}0365213 * V4 - 1{,}0380212 * V6 - 0{,}3593867 * V8 - -0{,}5890128 * V9 - 0{,}4062713 * V13 + 0{,}3142178 * V19 - 0{,}3140372 * V20 + 0{,}3521083 * V21 - 0{,}7804431 * V24 - 0{,}6703862 * V26 + 0{,}6607372 * V28 + 0{,}0012403 * Amount$$

Essa equação descreve o Intercepto (-4,3483670) que é o valor quando todas as variáveis independentes têm valor zero. Os valores apresentados à frente de cada variável são os coeficientes de cada variável explicativa. Ao final a composição do cálculo da contribuição de cada variável independente determinará a probabilidade de ocorrência do evento de fraude.

O Quadro 14 apresenta o resultado o cálculo de dois exemplos utilizando a equação anterior.

Quadro 14- Exemplo de cálculo para estimação da probabilidade de fraude e não fraude e classificação

Exemplo Definição Não Fraude			
Variáveis	Estimadores	Variáveis	Cálculo
(Intercept)	-4,3484	-	-4,3484
Time	0,0000	9,000	0,0000
V4	1,0365	2,820	2,9226
V6	-1,0380	0,070	-0,0722
V8	-0,3594	1,007	-0,3619
V9	-0,5890	0,000	0,0000
V13	-0,4063	-0,150	0,0611
V19	0,3142	0,304	0,0954
V20	-0,3140	-0,247	0,0776
V21	0,3521	-0,634	-0,2232
V24	-0,7804	0,256	-0,2000
V26	-0,6704	0,083	-0,0557
V28	0,6607	3,680	2,4315
Amount	0,0012	0,000	0,0000
		Calculo Predição	0,3268
		Classificação	Não Fraude

Exemplo Definição Fraude			
Variáveis	Estimadores	Variáveis	Cálculo
(Intercept)	-4,3484	-	-4,3484
Time	0,0000	339	0,0000
V4	1,0365	1,930	2,0006
V6	-1,0380	-0,861	0,8933
V8	-0,3594	-0,461	0,1655
V9	-0,5890	-0,646	0,3808
V13	-0,4063	-0,646	0,2626
V19	0,3142	0,492	0,1545
V20	-0,3140	-0,167	0,0524
V21	0,3521	0,680	0,2394
V24	-0,7804	-0,162	0,1266
V26	-0,6704	0,069	-0,0464
V28	0,6607	1,620	1,0704
Amount	0,0012	2,000	0,0025
		Calculo Predição	0,9537
		Classificação	Cálculo

Fonte: Elaboração Própria

Nos exemplos descritos no Quadro 14 é possível verificar:

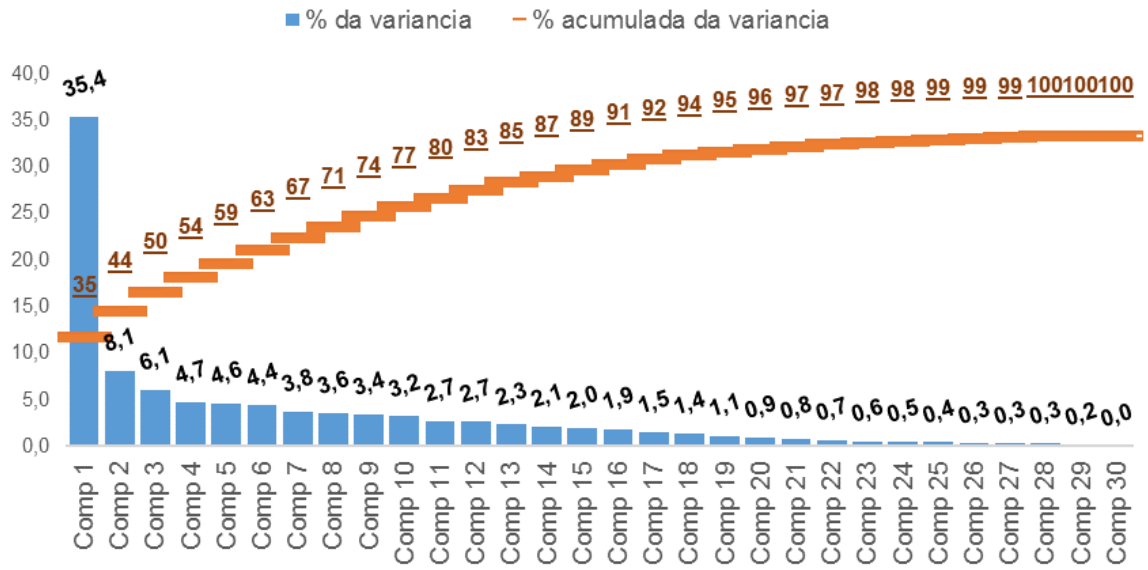
- O campo Variáveis que faz referência às variáveis utilizadas no modelo.
- As estimativas dos parâmetros do modelo conforme descrito no Quadro 12.
- As Variáveis são os valores da variável original obtido da base de predição dos dados.
- O Cálculo do valor da estimativa do parâmetro correspondente multiplicado pelo valor da Variável.
- Ao final tem-se o valor do Cálculo da Predição que é a soma dos valores da coluna Cálculo. Este valor é utilizado para a definição de qual grupo o registro será classificado, fraude ou não fraude. Tem-se por padrão a utilização do valor de 0,5 como referência para definição do grupo de classificação, contudo, esta definição depende do contexto específico do problema em análise.

9.6. Base de dados variáveis componentes transformadas (PCA)

Como forma adicional de resolver o problema será realizado o ajuste dos dados através da Análise de Componentes Principais (PCA) via Matriz de Correlação devido às características da base utilizada já descritas na seção 8 **(Silva,2008).**

Para aplicação do PCA utilizou-se do pacote *MVar.pt* versão 2.1.1 do *Software* R versão 4.2.3 (**OSSANI e CIRILO, 2023).** Os resultados obtidos são apresentados a seguir.

Gráfico 7- Os autovalores e percentuais de variância total explicada



Fonte: Elaboração Própria

O Gráfico 7 apresenta no eixo x a relação das componentes principais ordenadas pelo percentual de variância explicada, já no eixo y tem-se a variância explicada por cada componente. As barras horizontais em alaranjado apresentam a soma acumulada da variância das componentes. A análise do Gráfico 7 nos permite concluir que:

- A componente 1 responde por 35% da explicação da variância total. Este valor é maior que os valores das demais componentes que tiveram representatividade bem inferior.
- O conjunto de componentes de 2 a 30 apresentaram baixo valor individual de explicação da variância total, o que leva a utilização de um maior número de

componentes para que se tenha uma maior representatividade da variância total.

- Decide-se pela utilização de 18 componentes Comp. 1 a Comp. 18 que juntas representam uma variância total explicada de 94,68% da variância total que é de 30.

Definidas as 18 componentes como mais representativas passa-se a próxima etapa de construção do modelo logístico. As demais análises serão realizadas nas seções seguintes determinando a relevância de cada componente nos indicadores do modelo.

9.7. Modelo 2: regressão logística para variáveis transformadas PCA

Nesta etapa será desenvolvido um modelo de regressão logística sobre a base de dados já previamente tratada conforme descrito na seção 9.6 que apresenta as variáveis transformadas pela aplicação do PCA.

A base de dados possui 19 variáveis, descritas como Comp.1, Comp.2, Comp.3, Comp.4, Comp.5, Comp.6, Comp.7, Comp.8, Comp.9, Comp.10, Comp.11, Comp.12, Comp.13, Comp.14, Comp.15, Comp.16, Comp.17, Comp.18 e Class, sendo Class a variável resposta e as demais, variáveis explicativas. A base de dados foi separada em duas partes: base de teste e base de treino.

Foi ajustado o modelo logístico com todas as 18 componentes principais, mais a variável resposta obtendo-se o resultado do Quadro 15.

Quadro 15- Estimativas dos parâmetros do modelo logístico variáveis PCA com todas as variáveis

```
Call:
glm(formula = Class ~ ., family = binomial, data = treino_2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1638  -0.1996  -0.1005  -0.0361   3.4756

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.6772     0.3951  -1.714 0.086571 .
Comp.1        2.7615     0.3586   7.700 1.36e-14 ***
Comp.2        0.8129     0.2292   3.547 0.000389 ***
Comp.3        0.9195     0.2219   4.144 3.42e-05 ***
Comp.4       -1.4561     0.2051  -7.100 1.25e-12 ***
Comp.5        1.5200     0.2485   6.117 9.53e-10 ***
Comp.6       -0.4916     0.1763  -2.789 0.005285 **
Comp.7       -0.1849     0.1703  -1.086 0.277600
Comp.8        0.2999     0.1616   1.856 0.063425 .
Comp.9       -0.2232     0.1565  -1.426 0.153915
Comp.10      -0.4870     0.1784  -2.730 0.006325 **
Comp.11      -1.9433     0.3012  -6.453 1.10e-10 ***
Comp.12      -0.3704     0.2116  -1.750 0.080038 .
Comp.13      -0.8499     0.2104  -4.039 5.37e-05 ***
Comp.14       1.5016     0.2612   5.749 8.99e-09 ***
Comp.15       0.3612     0.2234   1.617 0.105948
Comp.16      -0.5571     0.2135  -2.610 0.009053 **
Comp.17      -0.4156     0.2726  -1.524 0.127456
Comp.18       0.7145     0.3000   2.382 0.017235 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1700.00  on 1666  degrees of freedom
Residual deviance:  314.65  on 1648  degrees of freedom
AIC: 352.65

Number of Fisher Scoring iterations: 10
```

Fonte: Elaboração Própria
“.” – Considerar ponto como separador de decimal.

Com exceção das componentes 7,8,9,12,15 e 17, todas as outras foram significativas ao nível de 5% de significância pois apresentaram p-value menores que 0,05.

O Null Deviance é de 1700,00 e o Residual Deviance 314,65. Tem-se ainda o valor do AIC de 352,65.

No Quadro 16 apresenta-se os resultados do ajuste do modelo retirando-se as componentes principais Comp.7, Comp.8, Comp.9, Comp.12, Comp.15 e Comp.17. Após o primeiro processamento, verifica-se que a componente principal Comp.16 torna-se estatisticamente não significativa com p-value de 0,07421 sendo também retirada (saída computacional não apresentada).

Quadro 16- Estimativas dos parâmetrosdo modelo logístico variáveis PCA otimizado.

```
Call:
glm(formula = Class ~ Comp.1 + Comp.2 + Comp.3 + Comp.4 + Comp.5
    Comp.6 + Comp.10 + Comp.11 + Comp.13 + Comp.14 + Comp.18,
    family = binomial, data = treino_2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4887  -0.2222  -0.1139  -0.0436   3.5332

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.3115     0.5222  -0.596 0.550892
Comp.1        2.9655     0.4674   6.345 2.23e-10 ***
Comp.2        0.8352     0.1893   4.411 1.03e-05 ***
Comp.3        0.8823     0.1808   4.879 1.07e-06 ***
Comp.4       -1.2913     0.1818  -7.102 1.23e-12 ***
Comp.5        1.2628     0.2052   6.155 7.50e-10 ***
Comp.6       -0.5974     0.1683  -3.549 0.000387 ***
Comp.10      -0.5108     0.1650  -3.095 0.001969 **
Comp.11      -1.5692     0.2633  -5.959 2.54e-09 ***
Comp.13      -0.8938     0.1916  -4.665 3.09e-06 ***
Comp.14       1.2472     0.2284   5.461 4.74e-08 ***
Comp.18       0.8449     0.2672   3.161 0.001570 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1700.00  on 1666  degrees of freedom
Residual deviance:  334.02  on 1655  degrees of freedom
AIC: 358.02

Number of Fisher Scoring iterations: 10
```

Fonte: Elaboração Própria
“.” – Considerar ponto como separador de decimal.

No Quadro 16, tem-se os resultados de Residual Deviance (334,02) e AIC (358,02), e observa-se que os valores encontrados após a retirada das componentes não significativas.

Os resultados da capacidade assertiva do modelo do Quadro 16 tanto na fase de treino quanto na de teste são apresentados no Quadro 17. Passa-se a predição da base de teste com apuração dos resultados.

9.8. Resultados do modelo de regressão logística para variáveis transformadas PCA

No Quadro 17 são apresentados os resultados obtidos no treino e no teste do modelo ajustado destacando-se a seguir:

- A Taxa de detecção de fraude no momento do Treino foi de 87,53% se confirmando na base de Teste com 88,43%. (Sensibilidade)

- A Especificidade no momento do Treino foi de 99,47% se confirmando na base de Teste com 99,29%.

A Taxa do Falso Alarme no momento do Treino foi de 0,53% passando para 0,70% na base de teste.

Quadro 17- Matriz confusão e resultados do modelo logístico com as variáveis PCA treino e teste

Matriz de Confusão -Base Treino

		Real		
		Não Fraude	Fraude	Tot.
Predito	Não Fraude	1315	43	1358
	Fraude	7	302	309
	Tot	1322	345	1667

Matriz de Confusão -Base Teste

		Real		
		Não Fraude	Fraude	Tot.
Predito	Não Fraude	563	17	580
	Fraude	4	130	134
	Tot	567	147	714

Proporções de Acertos		
	Não Fraude	Fraude
	0,99470	0,87536

Proporções de Acertos		
	Não Fraude	Fraude
	0,9929453	0,8843537

Proporções de Erros		
	Não Fraude	Fraude
	0,00530	0,12464

Proporções de Erros		
	Não Fraude	Fraude
	0,0070547	0,1156463

Proporções de Acertos totais
0,97001

Proporções de Erros totais
0,02999

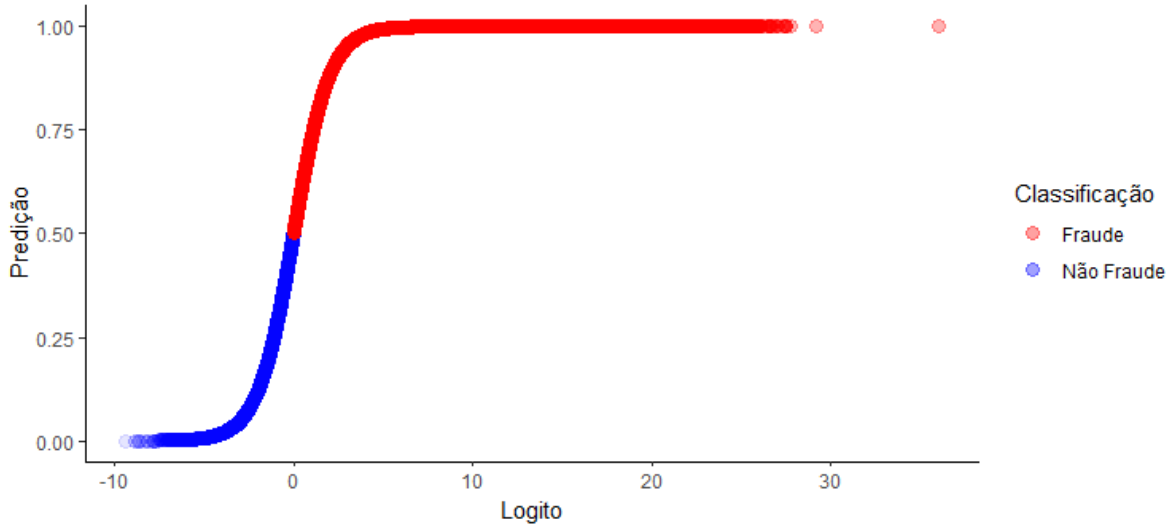
Proporções de Acertos totais
0,97059

Proporções de Erros totais
0,02941

Fonte: Elaboração Própria

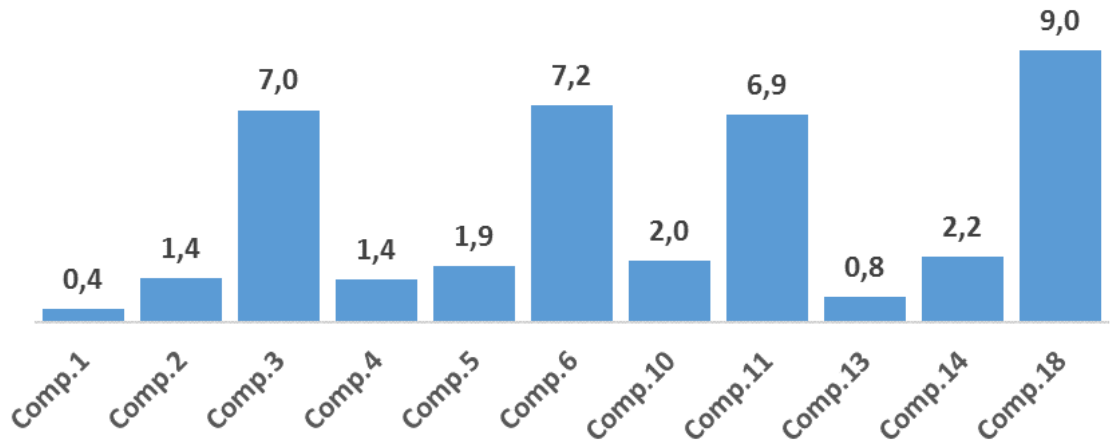
O Gráfico 8 demonstra o bom ajuste do modelo logístico, separando os registros de fraude e não fraude e atribuindo uma maior probabilidade de risco para os casos de fraude.

Gráfico 8 - Gráfico da curva logística PCA



Fonte: Elaboração Própria

Gráfico 9- Gráfico da Razão de Chances (Modelo Logístico PCA)



Fonte: Elaboração Própria

O Gráfico 9 demonstra a razão de chances para cada variável do modelo permitindo determinar a contribuição de cada variável a partir de sua ocorrência.

A equação do modelo logístico ajustado com as variáveis PCA (ver Quadro 16) é descrita como:

:

$$P(Y = 1) = -0{,}3115 + 2{,}9655 * Comp.1 + 0{,}8352 * Comp.2 + 0{,}8823 * Comp.3 - 1{,}2913 * Comp.4 + 1{,}2628 * Comp.5 - 0{,}5974 * Comp.6 - 0{,}5108 * Comp.10 - 1{,}5692 * Comp.11 - 0{,}8938 * Comp.13 + 1{,}2472 * Comp.14 + 0{,}8449 * Comp.18$$

O Quadro 18 apresenta o resultado o cálculo de dois exemplos utilizando a equação anterior.

Quadro 18 - Exemplo de cálculo para definição de fraude e não fraude Modelo Logístico PCA

Exemplo Definição Não Fraude			
Variáveis	Estimadores	Variáveis	Cálculo
(Intercept)	-0,3115	-	-0,3115
Comp.1	2,9655	1,609	4,7724
Comp.2	0,8352	-0,729	-0,6086
Comp.3	0,8823	-0,579	-0,5107
Comp.4	-1,2913	0,239	-0,3086
Comp.5	1,2628	0,644	0,8130
Comp.6	-0,5974	-0,308	0,1837
Comp.10	-0,5108	1,431	-0,7311
Comp.11	-1,5692	0,659	-1,0341
Comp.13	-0,8938	0,809	-0,7227
Comp.14	1,2472	-1,269	-1,5823
Comp.18	0,8449	0,159	0,1339
		Calculo Predição	0,0934
		Classificação	Não Fraude

Exemplo Definição Fraude			
Variáveis	Estimadores	Variáveis	Cálculo
(Intercept)	-0,3115	-	-0,3115
Comp.1	2,9655	0,508	1,5059
Comp.2	0,8352	-1,219	-1,0179
Comp.3	0,8823	1,5347012	1,3541
Comp.4	-1,2913	0,576	-0,7439
Comp.5	1,2628	2,079	2,6247
Comp.6	-0,5974	1,171	-0,6996
Comp.10	-0,5108	0,967	-0,4940
Comp.11	-1,5692	-0,210	0,3298
Comp.13	-0,8938	0,666	-0,5952
Comp.14	1,2472	0,8235282	1,0271
Comp.18	0,8449	-2,429	-2,0523
		Calculo Predição	0,9273
		Classificação	Fraude

Fonte: Elaboração Própria

No exemplo do Quadro 18 tem-se um caso de não fraude onde a probabilidade registrada foi de 0,0934 e um caso de fraude onde a probabilidade da existência do evento foi de 0,9273.

10. CONSIDERAÇÕES FINAIS

O presente trabalho inicia-se com uma exploração bibliográfica do conjunto de eventos históricos que contribuíram para a expansão do mercado de crédito. Dentre eles, cita-se a abertura comercial, a implantação do Plano Real e o tripé macroeconômico, que permitiram o controle da inflação, a redução dos juros e um maior controle da economia.

O texto contextualiza a expansão do volume de transações digitais e os desafios das instituições na busca de novos mecanismos de segurança e controle de fraudes. Foram relacionadas as principais modalidades de fraude cometidas atualmente, bem como seus impactos e formas de atuação, tendo destaque o *Worm* (atividade relacionada ao processo automatizado de propagação de códigos maliciosos na rede), o *Scan* (escaneamento de serviços ativos no computador com o intuito de identificar possíveis vulnerabilidades) e a fraude (qualquer ato ardiloso, enganoso, de má-fé, com o intuito de lesar ou ludibriar outrem ou de não cumprir determinado dever).

A proposta de criação de dois modelos, sendo o modelo logístico com as variáveis originais do banco de dados e o modelo com as componentes principais construídas via matriz de correlação dessas variáveis foi bem-sucedida**,** demonstrando que ambos os modelos apresentam bons ajustes para detecção de fraude.

Os resultados demonstraram que a aplicação do PCA para tratamento dos dados reduziu a quantidade de variáveis de 13 no modelo logístico com variáveis originais**,** para 11 na aplicação do PCA, permitindo uma melhor eficiência dos resultados do modelo de regressão logística. Contudo por se tratar de um modelo de aplicação teórica, recomenda-se novos estudos sob o tema que contribuam com esta constatação.

Por fim, conforme já descrito anteriormente, este trabalho tem caráter acadêmico utilizando dados que representam a realidade de instituições bancárias europeias. Desta forma os resultados não devem ser reproduzidos para o cenário brasileiro sem as devidas adaptações de acordo com as necessidades e contextos específicos.

Como sugestão para trabalhos futuros, propõe-se a aplicação de regras de cadeia de Markov **(MEDEIROS, SANTOS e OLIVEIRA, 2011),** para ampliação dos cenários, a fim de explorar as potencialidades da aplicação do modelo PCA juntamente com outras técnicas estatísticas.

REFERÊNCIAS

BACEN, **Fintechs de crédito e bancos digitais**, Estudo Especial nº 89/2020 – Divulgado originalmente como boxe do Relatório de Economia Bancária (2019). Disponível em: https://www.bcb.gov.br/conteudo/relatorioinflacao/EstudosEspeciais/EE089_Fintechs_de_credito_e_bancos_digitais.pdf. Acesso em: 15/01/2023

BASTOS, Paulo Sergio Siqueira; PEREIRA, Roberto Miguel. **Fraudes Eletrônicas: O que há de Novo?** 2007. Revista de Contabilidade do Mestrado em Ciências Contábeis UERJ, Rio de Janeiro, v.12 n.2. Disponível em: http://www.atena.org.br/revista/ojs-2.2.3-06/index.php/UERJ/article/view/639/635 Acesso em: 15/03/2023

BECKER, R. A., CHAMBERS, J. M. and WILKS, A. R. (1988) *The New S Language*. Wadsworth & Brooks/Cole. Disponível em: https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/sample

BRESLOW, NE; DAY, NE **Métodos estatísticos na pesquisa do câncer**. Volume II - o desenho e análise de estudos de coorte. IARC Sci Publ, v. 82, p. 1-406, 1987. Disponível em: https://www.iarc.who.int/publications/iarc-scientific-publications/stat/item/11023/Scientific-Publication-No-82-Statistical-Methods-in-Cancer-Research- Vol.-II-O-Desenho-e-Análise-de-Estudos-de-Coorte Acesso em: 15/04/2023

CERT, O Centro de Estudos, Resposta e Tratamento de Incidentes de Segurança no Brasil **Estatística dos Incidentes Reportados Ao CERT.br** 2022. Disponível em: https://www.cert.br/stats/incidentes/

CYSNE, Rubens Penha, COSTA Sergio Gustavo Silveira da **Reflexo do Plano Real sobre o Sistema Bancário Brasileiro,** 1996 Disponível em: https://bibliotecadigital.fgv.br/dspace/bitstream/handle/10438/931/000065292.pdf?sequence=1&isAllowed=y Acesso em: 15/04/2023

EBC, Empresa Brasil de Comunicação. **Cartões lideram forma de pagamento em 2021 com participação de 51%** 2022. Disponível em: https://agenciabrasil.ebc.com.br/economia/noticia/2022-11/cartoes-lideraram-forma-de-pagamento-em-2021-com-participacao-de-51Acesso em: 25/03/2023

EMILIANO, Paulo Cezar; **Fundamentos e Aplicações dos Critérios de Informação: Akaike e Bayesiano**, 2009. Disponível em: http://repositorio.ufla.br/jspui/bitstream/1/3636/1/DISSERTA%C3%87%C3%83O_Fundamentos%20e%20Aplica%C3%A7%C3%B5es%20dos%20Crit%C3%A9rios%20de%20Informa%C3%A7%C3%A3o%20Akaike%20e%20Bayesiano.pdf Acesso em: 18/06/2023

FARIAS, Daniel Teixeira de. **Análise da capacidade preditiva de modelos de regularização aplicado ao IPCA**. 2021. Disponível em: https://repositorio.ufc.br/handle/riufc/59113

FEBRABAN, **Investimentos de bancos com tecnologia crescem 48% em 2019 e orçamento total chega a R$ 24,6 bilhões**, 2020 Disponível em: https://portal.febraban.org.br/noticia/3470/pt-br/

FEEB-PR - A Federação dos Empregados em Estabelecimentos Bancários no Estado do Paraná **UMA BREVE HISTÓRIA DA DIGITALIZAÇÃO DOS BANCOS NO BRASIL,** 2021.
Disponível em: https://www.bancariosparanagua.org.br/noticia/uma-breve-historia-da-digitalizacao-dos-bancos-no-brasilAcesso em: 15/01/2023

FRIEDMAN, Jerome, HASTIE, Trevor e TIBSHIRANI, Rob. **Regularization Paths for Generalized Linear Models via Coordinate Descent**. Journal of Statistical Software, v. 33, n. 1, pág. 1-22, fev. 2010. Disponível em: https://www.jstatsoft.org/article/view/v033i01 Acesso em: 05/05/2023

FRIEDMAN, Jerome, HASTIE, Trevor e TIBSHIRANI, Rob.**glmnet: Modelos lineares generalizados regularizados de laço e rede elástica.** Versão do pacote R 4.1-1. 2020. Obtido em https://CRAN.R-project.org/package=glmnet Acesso em: 05/05/2023

HONGYU, Kuang; SANDANIELO, Vera Lúcia Martins; JUNIOR, Gilmar Jorge de Oliveira. **Análise de Componentes principais: resumo teórico, aplicação e interpretação**. 2015, Disponível em: https://periodicoscientificos.ufmt.br/ojs/index.php/eng/article/view/3398

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). ***Applied Logistic Regression* (3rd ed.)**. John Wiley & Sons.

KAGGLE, **Machine Learning para Competições Kaggle. 2010** Disponível em: https://www.kaggle.com/ Acesso em: 19/03/2020.

KUHN, Max. **Variable selection using the caret package**. **2010**. Disponível em: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=56f21af0f48d8f9119eac0cc8c7082e154427705

LEBICHOT, BERTRAND; VERHELST, THEO; BORGNE, YANN-A"EL LE; HE-GUELTON, LIYUN; OBLÉ, FRÉDÉRIC; BONTEMPI, GIANLUCA, "**Transfer Learning Strategies for Credit Card Fraud Detection**", 2021 In: IEEE access, vol. 9, pp. 114754-114766, (DOI: 10.1109/ACCESS.2021.3104472).Disponível em: https://ieeexplore.ieee.org/document/9512084Acesso em: 15/04/2023

MACIEL, Raul Lucas Tanigut Brisola **Profissional 4.0: perspectivas para formação e atuação dos profissionais de contabilidade e finanças na Economia 4.0 - Breve histórico da regulação dos bancos digitais no Brasil** 2018. VIII SIMPÓSIO DE CONTABILIDADE E FINANÇAS DE DOURADOS – SICONF. Disponível em:

https://d1wqtxts1xzle7.cloudfront.net/61840563/Breve_historico_da_regulacao_dos_bancos_digitais_no_Brasil20200120-115309-1knazvu-libre.pdf?1579551350=&response-content-disposition=inline%3B+filename%3DBreve_historico_da_regulacao_dos_bancos.pdf&Expires=1673825683&Signature=PMptHDw7wjLEETD4BmjtVYBPhXIuwKJzNmvYmcd1X0u9~7q3L8VeFxtilhw-Cclu0J65IwOxMEjgpYYwZWLKJYU8OBxcFA88UdOOZcvVptr8LO8kZhC4iYjh63uw7wJ3oTsQ87n66MrpEQRHyabZqVl2Cr6dD4FPJ6IaYmguo3Yn8~NQlj6GzhuxP9Ru1aGcxfLWrGM~kycp3QuCBAXP~8Hh7NxZD0n8CDhltYEGUMlxnrCwzEUrt3on7-F-7EI9IMvd-~5C9p2y0~WONPTQ2cP7EXmpL4Ez-CDIO2EYIUEsDjfeADX48BCELVAIUFHWaVzpyQypSbbaSxJYTZwe4g__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZAAcesso em: 05/05/2023

MEDEIROS, Rubem José Vasconcelos; SANTOS, Raphael Sousa; OLIVEIRA, Liduino José Pitombeira. **Modelagem Sistêmica baseada em Cadeias de Markov**. 2011. Universidade Federal de Campina Grande. Disponível em: https://d1wqtxts1xzle7.cloudfront.net/37214899/MODELAGEM_SISTEMICA_BASEADA_EM_CADEIAS_DE_MARKOV-libre.pdf?1428249630=&response-content-disposition=inline%3B+filename%3DModelagem_Sistemica_Baseada_em_Cadeias_d.pdf&Expires=1711920483&Signature=TKKWlnqvHXvScpYRoTQfejx5loGMpjhmlZ4VvqxMOzDCkQBAMTiDG1dLKdnpLzlKSWviQQdO907xOVL8wc0mku-qWmgNEwbhd0En11m4ezjKgU-WRGmYnzE2FTWIZO34HK6Dg6XmyufFjxRf0Dfz~HgH2C4sc3SgmnLUmxbMfHJlDcOipumw0Trff2PskSo8FJwSi3l6O1Jb6AcHj-OrmF0MVISZWa6dh3H2P5c8MDZjmLBIZUbzgKO9jJ~2iPVNGKwVSN1CNbXpuiN3d0ZygM6HqTMVlBKf90RGTG2W7181FPWH1UYI2eb44FQ1xD0N8y2iifnVUyUX7ZmcN3ebqg__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA. Acesso em: 31/03/2024

MINGOTI, Sueli Aparecida, **Análise de Dados através de Métodos de Estatística Multivariada,** 2013. Editora UFMG.

MOHANTY, Soumya Ranjan; **MAHAPATRA**, Ajit Kumar. **Multicollinearity: Effects, Symptoms, and Solutions. Journal of Statistics Applications & Probability**, v. 7, n. 2, p. 355-366, 2018. Disponível em: https://www.researchgate.net/publication/327370167_Multicollinearity_Effects_Symptoms_and_Solutions. Acesso em: 15 abr. 2023.

MORA. Mônica e IPEA, **A EVOLUÇÃO DO CRÉDITO NO BRASIL ENTRE 2003 E 2010** 2015[online]. Rio de Janeiro, janeiro de 2015 IPEA. Disponível em: http://repositorio.ipea.gov.br/handle/11058/3537Acesso em: 11/12/2022

NAKAMURA, Karina Gernhardt **Multicolinearidade em modelos de regressão logística** 2013 Universidade de São Paulo. Dissertação Mestrado de Ciências. Disponível em: https://www.teses.usp.br/teses/disponiveis/45/45133/tde-28052013-222241/publico/Dissertacao_KarinaGernhardtNakamura.pdfAcesso em: 05/05/2023

OSSANI, Paulo Cesar; CIRILLO, Marcelo Angelo. **MVar.pt: Análise multivariada (brazilian portugueses) version 2.2.1.** 2023. Disponível em: https://cran.rstudio.com/web/packages/MVar.pt/index.html. Acesso em: 16 abr. 2023.

PRIPERAM, Informática S.A., **Dicionario.priberam.org.** 2022.Disponível em: https://www.dicionario.priberam.org/fraude. Acesso em: 11/11/2022

SANTOS, Artur Trazola – **Abertura comercial na década de 1990 e os impactos na indústria automobilística** – Periódico PUC MINAS - Revista Fronteira Belo Horizonte v.8 n.16 2º sem. 2009 Belo Horizonte; 2009. Disponível em: https://periodicos.pucminas.br/index.php/fronteira/article/view/3860/4160. Acesso em: 31 ago. 2022.

SILVA, Jonas Guedes Borges da. **Aplicação da Análise de componentes Principais (PCA) no diagnóstico de defeitos de rolamentos através da assinatura elétrica de motores de indução.** 2008 UNIFEI – Dissertação Mestrado em Ciências e Engenharia Elétrica. Disponível em: https://repositorio.unifei.edu.br/xmlui/bitstream/handle/123456789/1665/dissertacao_0032606.pdf?sequence=1&isAllowed=y Acesso em: 24/04/2023

SILVA, Roseli Basso, Grupo A – **Inflação e Taxa de Juros no Brasil (Plano real até 2002)** 2015. Disponível em: https://randomwalk.com.br/tag/plano-real/#:~:text=Focando%20o%20objetivo%20de%20estabilidade,monet%C3%A1ria%20r%C3%ADgida%20(juros%20altos). Acesso em: 02/09/2022.

VARELLA, Carlos Alberto Alves, **Análise Multivariada Aplicada as Ciências Agrárias – Análise de Componentes Principais** Universidade Federal Rural do Rio de Janeiro – Pós-Graduação em Agronomia. 2008. Disponível em: http://www.ufrrj.br/institutos/it/deng/varella/Downloads/multivariada%20aplicada%20as%20ciencias%20agrarias/Aulas/analise%20de%20componentes%20principais.pdf Acesso em: 19/08/2023

YAN, Weizhong; YE, Yangdong; CHEN, Philip. **Tamanhos de amostra ideais para classificar dados desequilibrados usando redes neurais**. Sistemas Baseados em Conhecimento, v. 113, p. 1-9, 2016. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0950705116301156 Acesso em: 05/05/2023

ANEXOS

ANEXO 1 – Seleção de amostra dos dados.

```
require("pak")
pak::pak("tidyverse/ggplot2")

library("caret")
library("remotes")
library("ggplot2")
library("plyr")
require("MASS")
require("exploreR")
require("readr")
require("skimr")
library("openxlsx")


# Leitura da Base de Dados

df_fraud_card_geral <-
read.csv("..\\TCC_Trabalho_Final_de_Estatistica\\Script_tcc_python_R\\creditcard.csv")
nrow(df_fraud_card_geral)

# Fazendo a seleção da amostra para tratamento

# Separando base fraude e não fraude

df_fraude <- df_fraud_card_geral[df_fraud_card_geral$Class == 1, ]
df_nao_fraude <- df_fraud_card_geral[df_fraud_card_geral$Class == 0, ]
nrow(df_fraude)
nrow(df_nao_fraude)

# define o número de registros que deseja selecionar aleatoriamente
num_registros <- 1889

# seleciona aleatoriamente os índices dos registros
```

```
indices_amostra <- sample(1:nrow(df_nao_fraude), num_registros, replace = FALSE)

# extrai os registros selecionados aleatoriamente
df_nao_fraude_amostra <- df_nao_fraude[indices_amostra, ]
nrow(df_nao_fraude_amostra)

df_fraud_card <- rbind(df_fraude, df_nao_fraude_amostra)
head(df_fraud_card)
nrow(df_fraud_card)
summary(df_fraud_card)
num_na_por_coluna <- colSums(is.na(df_fraud_card)) # verifica se existe valores vazios nas colunas do data frame
print(num_na_por_coluna)

# Salvando as bases de referencia do trabalho final

setwd("../TCC_Trabalho_Final_de_Estatistica/")

write.xlsx(df_fraud_card,"Base_amostra_balanceada_sample_2381.xlsx")
```

Fonte: Elaborado pelo autor, 2023

ANEXO 2 – Análise descritiva dos dados

```
# Analise Descritiva dos dados
# Sumarização Base total

df1 <- df_fraud_card_geral[df_fraud_card_geral$Class == 1, ]
df0 <- df_fraud_card_geral[df_fraud_card_geral$Class == 0, ]

summary0 <- summary(df0)
summary1 <- summary(df1)

summary1

summary0_std_dev <- data.frame(sapply(df0, function(x) if(is.numeric(x)) sd(x) else NA))
colnames(summary0_std_dev) <- "Desvio_Padrão"
summary0_std_dev$Desvio_Padrão <- sprintf("%.4f", summary0_std_dev$Desvio_Padrão)

summary1_std_dev <- data.frame(sapply(df1, function(x) if(is.numeric(x)) sd(x) else NA))
colnames(summary1_std_dev) <- "Desvio_Padrão"
summary1_std_dev$Desvio_Padrão <- sprintf("%.4f", summary1_std_dev$Desvio_Padrão)
```

```
summary0_std_dev
summary1_std_dev


# Exibição do novo data frame com o desvio padrão de cada coluna numérica


df0_g <- subset(df_fraud_card_geral, Class == 0)
df1_g <- subset(df_fraud_card_geral, Class == 1)

summary0_g <- summary(df0_g)
summary1_g <- summary(df1_g)

summary0_std_dev_g <- data.frame(sapply(df_fraud_card_geral, function(x) if(is.numeric(x)) sd(x) else NA))
names(summary0_std_dev) <- paste0("desvio_padrao_", names(std_dev))

summary1_std_dev_g <- data.frame(sapply(df_fraud_card_geral, function(x) if(is.numeric(x)) sd(x) else NA))
names(summary1_std_dev) <- paste0("desvio_padrao_", names(std_dev))

summary0_std_dev_g

setwd("../TCC_Trabalho_Final_de_Estatistica/")

write.xlsx(summary0,"summary0.xlsx")
write.xlsx(summary1,"summary1.xlsx")


# exibe os resultados dos summaries para cada subconjunto
print(summary1_std_dev_g)
print(summary1)

# Definir função personalizada para calcular moda, desvio padrão e média
my_summary <- function(x) {
  return(c(Média = mean(x), Desvio_Padrão = sd(x), Moda = names(sort(-table(x)))[1]))
}

# Aplicar a função summary com a função personalizada
summary0_compl <- summary(summary0, statistics = my_summary)

df_fraud_card$Class <- as.factor(df_fraud_card$Class)

# Analise Descritiva

df_time <- df_fraud_card_geral[,c(31,1)]
df_V1 <- df_fraud_card_geral[,c(31,2)]
df_V2 <- df_fraud_card_geral[,c(31,3)]
df_V3 <- df_fraud_card_geral[,c(31,4)]
```

```
df_V4 <- df_fraud_card_geral[,c(31,5)]
df_V5 <- df_fraud_card_geral[,c(31,6)]
df_V6 <- df_fraud_card_geral[,c(31,7)]
df_V7 <- df_fraud_card_geral[,c(31,8)]
df_V8 <- df_fraud_card_geral[,c(31,9)]
df_V9 <- df_fraud_card_geral[,c(31,10)]
df_V10 <- df_fraud_card_geral[,c(31,11)]
df_V11 <- df_fraud_card_geral[,c(31,12)]
df_V12 <- df_fraud_card_geral[,c(31,13)]
df_V13 <- df_fraud_card_geral[,c(31,14)]
df_V14 <- df_fraud_card_geral[,c(31,15)]
df_V15 <- df_fraud_card_geral[,c(31,16)]
df_V16 <- df_fraud_card_geral[,c(31,17)]
df_V17 <- df_fraud_card_geral[,c(31,18)]
df_V18 <- df_fraud_card_geral[,c(31,19)]
df_V19 <- df_fraud_card_geral[,c(31,20)]
df_V20 <- df_fraud_card_geral[,c(31,21)]
df_V21 <- df_fraud_card_geral[,c(31,22)]
df_V22 <- df_fraud_card_geral[,c(31,23)]
df_V23 <- df_fraud_card_geral[,c(31,24)]
df_V24 <- df_fraud_card_geral[,c(31,25)]
df_V25 <- df_fraud_card_geral[,c(31,26)]
df_V26 <- df_fraud_card_geral[,c(31,27)]
df_V27 <- df_fraud_card_geral[,c(31,28)]
df_V28 <- df_fraud_card_geral[,c(31,29)]
df_Amount <- df_fraud_card_geral[,c(31,30)]

par(mfrow= c(2,3))
plot(df_time, names=c("Não Fraude","Fraude"), cex=1.6)
plot(df_V1, names=c("Não Fraude","Fraude"), cex=1.6)
plot(df_V2, names=c("Não Fraude","Fraude"), cex=1.6)
plot(df_V3, names=c("Não Fraude","Fraude"), cex=1.6)
plot(df_V4, names=c("Não Fraude","Fraude"), cex=1.6)
plot(df_V5, names=c("Não Fraude","Fraude"), cex=1.6)

par(mfrow= c(2,3))
plot(df_V6, names=c("Não Fraude","Fraude"), cex=1.6)
plot(df_V7, names=c("Não Fraude","Fraude"), cex=1.6)
plot(df_V8, names=c("Não Fraude","Fraude"), cex=1.6)
plot(df_V9, names=c("Não Fraude","Fraude"), cex=1.6)
plot(df_V10, names=c("Não Fraude","Fraude"), cex=1.6)
plot(df_V11, names=c("Não Fraude","Fraude"), cex=1.6)

par(mfrow= c(2,3))
plot(df_V12, names=c("Não Fraude","Fraude"), cex=1.6)
plot(df_V13, names=c("Não Fraude","Fraude"), cex=1.6)
plot(df_V14, names=c("Não Fraude","Fraude"), cex=1.6)
plot(df_V15, names=c("Não Fraude","Fraude"), cex=1.6)
plot(df_V16, names=c("Não Fraude","Fraude"), cex=1.6)
plot(df_V17, names=c("Não Fraude","Fraude"), cex=1.6)
```

```
par(mfrow= c(2,3))
plot(df_V18, names=c("Não Fraude","Fraude"), cex=1.6)
plot(df_V19, names=c("Não Fraude","Fraude"), cex=1.6)
plot(df_V20, names=c("Não Fraude","Fraude"), cex=1.6)
plot(df_V21, names=c("Não Fraude","Fraude"), cex=1.6)
plot(df_V22, names=c("Não Fraude","Fraude"), cex=1.6)
plot(df_V23, names=c("Não Fraude","Fraude"), cex=1.6)

par(mfrow= c(2,3))
plot(df_V24, names=c("Não Fraude","Fraude"), cex=1.6)
plot(df_V25, names=c("Não Fraude","Fraude"), cex=1.6)
plot(df_V26, names=c("Não Fraude","Fraude"), cex=1.6)
plot(df_V27, names=c("Não Fraude","Fraude"), cex=1.6)
plot(df_V28, names=c("Não Fraude","Fraude"), cex=1.6)
plot(df_Amount, names=c("Não Fraude","Fraude"), cex=1.6)

head(df_time)

# Histograma Correlação

par(mfrow= c(2,3))

plot(density(df_time[df_time$Class=='0',]$Time), col="green", main="Densidade
TimexClass", cex=1.8)
polygon(density(df_time[df_time$Class=='0',]$Time), col="green")
lines(density(df_time[df_time$Class=='1',]$Time), col="red")

plot(density(df_V1[df_V1$Class=='0',]$V1), col="green", main="Densidade
V1xClass", cex=1.8)
polygon(density(df_V1[df_V1$Class=='0',]$V1), col="green")
lines(density(df_V1[df_V1$Class=='1',]$V1), col="red")

plot(density(df_V2[df_V2$Class=='0',]$V2), col="green", main="Densidade
V2xClass", cex=1.8)
polygon(density(df_V2[df_V2$Class=='0',]$V2), col="green")
lines(density(df_V2[df_V2$Class=='1',]$V2), col="red")

plot(density(df_V3[df_V3$Class=='0',]$V3), col="green", main="Densidade
V3xClass", cex=1.8)
polygon(density(df_V3[df_V3$Class=='0',]$V3), col="green")
lines(density(df_V3[df_V3$Class=='1',]$V3), col="red")

plot(density(df_V4[df_V4$Class=='0',]$V4), col="green", main="Densidade
V4xClass", cex=1.8)
polygon(density(df_V4[df_V4$Class=='0',]$V4), col="green")
lines(density(df_V4[df_V4$Class=='1',]$V4), col="red")

plot(density(df_V5[df_V5$Class=='0',]$V5), col="green", main="Densidade
V5xClass", cex=1.8)
```

```
polygon(density(df_V5[df_V5$Class=='0',]$V5), col="green")
lines(density(df_V5[df_V5$Class=='1',]$V5), col="red")

par(mfrow= c(2,3))

plot(density(df_V6[df_V6$Class=='0',]$V6),          col="green",          main="Densidade
V6xClass", cex=1.8)
polygon(density(df_V6[df_V6$Class=='0',]$V6), col="green")
lines(density(df_V6[df_V6$Class=='1',]$V6), col="red")

plot(density(df_V7[df_V7$Class=='0',]$V7),          col="green",          main="Densidade
V7xClass", cex=1.8)
polygon(density(df_V7[df_V7$Class=='0',]$V7), col="green")
lines(density(df_V7[df_V7$Class=='1',]$V7), col="red")

plot(density(df_V8[df_V8$Class=='0',]$V8),          col="green",          main="Densidade
V8xClass", cex=1.8)
polygon(density(df_V8[df_V8$Class=='0',]$V8), col="green")
lines(density(df_V8[df_V8$Class=='1',]$V8), col="red")

plot(density(df_V9[df_V9$Class=='0',]$V9),          col="green",          main="Densidade
V9xClass", cex=1.8)
polygon(density(df_V9[df_V9$Class=='0',]$V9), col="green")
lines(density(df_V9[df_V9$Class=='1',]$V9), col="red")

plot(density(df_V10[df_V10$Class=='0',]$V10),       col="green",       main="Densidade
V10xClass", cex=1.8)
polygon(density(df_V10[df_V10$Class=='0',]$V10), col="green")
lines(density(df_V10[df_V10$Class=='1',]$V10), col="red")

plot(density(df_V11[df_V11$Class=='0',]$V11),       col="green",       main="Densidade
V11xClass", cex=1.8)
polygon(density(df_V11[df_V11$Class=='0',]$V11), col="green")
lines(density(df_V11[df_V11$Class=='1',]$V11), col="red")

par(mfrow= c(2,3))

plot(density(df_V12[df_V12$Class=='0',]$V12),       col="green",       main="Densidade
V12xClass", cex=1.8)
polygon(density(df_V12[df_V12$Class=='0',]$V12), col="green")
lines(density(df_V12[df_V12$Class=='1',]$V12), col="red")

plot(density(df_V13[df_V13$Class=='0',]$V13),       col="green",       main="Densidade
V13xClass", cex=1.8)
polygon(density(df_V13[df_V13$Class=='0',]$V13), col="green")
lines(density(df_V13[df_V13$Class=='1',]$V13), col="red")

plot(density(df_V14[df_V14$Class=='0',]$V14),       col="green",       main="Densidade
V14xClass", cex=1.8)
polygon(density(df_V14[df_V14$Class=='0',]$V14), col="green")
```

```
lines(density(df_V14[df_V14$Class=='1',]$V14), col="red")

plot(density(df_V15[df_V15$Class=='0',]$V15),      col="green",      main="Densidade
V15xClass", cex=1.8)
polygon(density(df_V15[df_V15$Class=='0',]$V15), col="green")
lines(density(df_V15[df_V15$Class=='1',]$V15), col="red")

plot(density(df_V16[df_V16$Class=='0',]$V16),      col="green",      main="Densidade
V16xClass", cex=1.8)
polygon(density(df_V16[df_V16$Class=='0',]$V16), col="green")
lines(density(df_V16[df_V16$Class=='1',]$V16), col="red")

plot(density(df_V17[df_V17$Class=='0',]$V17),      col="green",      main="Densidade
V17xClass", cex=1.8)
polygon(density(df_V17[df_V17$Class=='0',]$V17), col="green")
lines(density(df_V17[df_V17$Class=='1',]$V17), col="red")


par(mfrow= c(2,3))

plot(density(df_V18[df_V18$Class=='0',]$V18),      col="green",      main="Densidade
V18xClass", cex=1.8)
polygon(density(df_V18[df_V18$Class=='0',]$V18), col="green")
lines(density(df_V18[df_V18$Class=='1',]$V18), col="red")

plot(density(df_V19[df_V19$Class=='0',]$V19),      col="green",      main="Densidade
V19xClass", cex=1.8)
polygon(density(df_V19[df_V19$Class=='0',]$V19), col="green")
lines(density(df_V19[df_V19$Class=='1',]$V19), col="red")

plot(density(df_V20[df_V20$Class=='0',]$V20),      col="green",      main="Densidade
V20xClass", cex=1.8)
polygon(density(df_V20[df_V20$Class=='0',]$V20), col="green")
lines(density(df_V20[df_V20$Class=='1',]$V20), col="red")

plot(density(df_V21[df_V21$Class=='0',]$V21),      col="green",      main="Densidade
V21xClass", cex=1.8)
polygon(density(df_V21[df_V21$Class=='0',]$V21), col="green")
lines(density(df_V21[df_V21$Class=='1',]$V21), col="red")

plot(density(df_V22[df_V22$Class=='0',]$V22),      col="green",      main="Densidade
V22xClass", cex=1.8)
polygon(density(df_V22[df_V22$Class=='0',]$V22), col="green")
lines(density(df_V22[df_V22$Class=='1',]$V22), col="red")

plot(density(df_V23[df_V23$Class=='0',]$V23),      col="green",      main="Densidade
V23xClass", cex=1.8)
polygon(density(df_V23[df_V23$Class=='0',]$V23), col="green")
lines(density(df_V23[df_V23$Class=='1',]$V23), col="red")
```

```
par(mfrow= c(2,3))

plot(density(df_V24[df_V24$Class=='0',]$V24), col="green", main="Densidade
V24xClass", cex=1.8)
polygon(density(df_V24[df_V24$Class=='0',]$V24), col="green")
lines(density(df_V24[df_V24$Class=='1',]$V24), col="red")

plot(density(df_V25[df_V25$Class=='0',]$V25), col="green", main="Densidade
V25xClass", cex=1.8)
polygon(density(df_V25[df_V25$Class=='0',]$V25), col="green")
lines(density(df_V25[df_V25$Class=='1',]$V25), col="red")

plot(density(df_V26[df_V26$Class=='0',]$V26), col="green", main="Densidade
V26xClass", cex=1.8)
polygon(density(df_V26[df_V26$Class=='0',]$V26), col="green")
lines(density(df_V26[df_V26$Class=='1',]$V26), col="red")

plot(density(df_V27[df_V27$Class=='0',]$V27), col="green", main="Densidade
V27xClass", cex=1.8)
polygon(density(df_V27[df_V27$Class=='0',]$V27), col="green")
lines(density(df_V27[df_V27$Class=='1',]$V27), col="red")

plot(density(df_V28[df_V28$Class=='0',]$V28), col="green", main="Densidade
V28xClass", cex=1.8)
polygon(density(df_V28[df_V28$Class=='0',]$V28), col="green")
lines(density(df_V28[df_V28$Class=='1',]$V28), col="red")

plot(density(df_Amount[df_Amount$Class=='0',]$Amount), col="green",
main="Densidade AmountxClass", cex=1.8)
polygon(density(df_Amount[df_Amount$Class=='0',]$Amount), col="green")
lines(density(df_Amount[df_Amount$Class=='1',]$Amount), col="red")
polygon(density(df_Amount[df_Amount$Class=='1',]$Amount), col="red")

# Analise de correlação

require(corrplot)
library(RColorBrewer)

par(mfrow=c(1,1))

# Define as paletas de cores
my_colors <-
c("#DCDCDC","#D1D1D1","#C6C6C6","#BDBDBD","#B2B2B2","#A9A9A9")
my_colors_above <- c( "#483D8B",
"#00008B","#363636","#1C1C1C","#000000","#000000")

# Define a função para a escala de cores
my_color_fun <- colorRampPalette(colors = c(my_colors, my_colors_above))

df_fraud_card$Class <- as.numeric(as.character(df_fraud_card$Class))
```

```
# Plota a matriz de correlação com os valores ajustados para a escala de cores
par(mfrow=c(1,1))
corrplot(cor(df_fraud_card[,c(1,2,5,6,7,9,10,11,12,13,14,15,16,20,22,23,24,25,26,27,
28,29,30)]),
        method = "number", type = "lower", upper = "circle",
        col = my_color_fun(100), number.color = "black", bg = "white",
        number.cex = 0.8, addCoef.col = "black", cl.pos = "n",
        cl.length = 1, cl.lim = c(0.3,1))

par(mfrow=c(1,1))
corrplot(cor(df_fraud_card_geral[,c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,2
0,21,22,23,24,25,26,27,28,29,30)]),
        method = "number", type = "lower", upper = "circle",
        col = my_color_fun(100), number.color = "black", bg = "white",
        number.cex = 0.5, addCoef.col = "black", cl.pos = "n",
        cl.length = 1, cl.lim = c(0.3,1))

cor_df                                                                    <-
cor(df_fraud_card_geral[,c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,
23,24,25,26,27,28,29,30)])
round(cor_df, 4)

df_fraud_card

cor_df_amostra                                                            <-
cor(df_fraud_card[,c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,
25,26,27,28,29,30)])
round(cor_df_amostra, 4)

df_fraud_card$Class <- as.numeric(as.character(df_fraud_card$Class))
par(mfrow=c(2,3))
corrplot(cor(df_fraud_card[,c(1,2,3,4,5,31)]), method="number")
corrplot(cor(df_fraud_card[,c(6,7,8,9,10,31)]), method="number")
corrplot(cor(df_fraud_card[,c(11,12,13,14,15,31)]), method="number")
corrplot(cor(df_fraud_card[,c(16,17,18,19,20,31)]), method="number")
corrplot(cor(df_fraud_card[,c(21,22,23,24,25,31)]), method="number")
corrplot(cor(df_fraud_card[,c(26,27,28,29,30,31)]), method="number")

df_fraud_card$Class <- as.factor(df_fraud_card$Class)

# Analise Univariada

require(exploreR)
require(knitr)

saida <- masslm(df_fraud_card_antiga, dv.var = "Class")
saida <- saida[order(saida$R.squared, decreasing=TRUE),]
names(saida) <- c("Variavel", "Coeficiente", "p-valor", "R2")
saida
```

```
knitr::kable(saida, digits = 4, row.names = FALSE,format.args = list(scientific = FALSE))

library("glmnet")
library("pors")

packageVersion("glmnet")

cvfit <- cv.glmnet(as.matrix(df_fraud_card[, -31]), df_fraud_card$Class, family = "binomial", alpha = 1)
coef(cvfit, s = "lambda.min")

print(cvfit$lambda)


cor_df_amostra <- cor(df_fraud_card[,c(1,2,5,6,7,9,10,11,12,13,14,15,16,20,22,23,24,25,26,27,28,29,30)])
round(cor_df_amostra, 4)


# Calcular os PORs
por <- pors(cvfit, x = x.train, y = y.train, method = "glmnet")

# Exibir os PORs
print(por$POR)

df_fraud_card[, -31]
```

Fonte: Elaborado pelo autor, 2023

ANEXO 3 – Matriz das correlações das variáveis com as componentes principais

	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20	V21	V22	V23	V24	V25	V26	V27	V28	Amount		
	7	4	7	3	3	5	9	5	6	3	3	4	3	4	4	7	3	4	5	7	9	5	6	6	7	9	30	7	8	6		
	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20	V21	V22	V23	V24	V25	V26	V27	V28	Amount		
Comp 1	-0,3	-0,7	0,7	-0,8	0,8	-0,7	-0,4	-0,7	0,1	-0,7	-0,8	0,8	-0,8	0,1	-0,7	-0,1	-0,8	-0,8	-0,7	0,4	0,3	0,0	0,1	0,0	0,0	0,0	9,5	0,0	-0,1	0,0	19	Comp 1
Comp 2	-0,2	0,2	-0,1	0,1	0,0	0,2	-0,3	0,2	0,2	0,2	0,1	0,3	-0,1	0,1	-0,4	0,3	-0,1	-0,1	0,0	0,1	-0,5	0,6	-0,6	-0,2	-0,2	0,2	1,9	0,6	0,3	0,0	14	Comp 2
Comp 3	0,0	0,2	-0,3	0,1	0,0	0,1	-0,5	0,2	0,7	0,0	0,0	0,0	-0,1	0,3	0,0	0,2	0,0	-0,1	0,0	0,2	0,1	-0,5	0,4	-0,5	0,1	-0,1	9,4	0,1	-0,2	0,1	9	Comp 3
Comp 4	0,0	0,1	0,1	-0,1	0,2	0,2	-0,2	0,1	0,0	-0,1	-0,1	0,2	-0,1	-0,2	-0,2	-0,3	0,2	0,2	0,3	-0,4	0,3	-0,2	0,1	0,1	-0,5	0,3	3,7	0,2	0,1	-0,2	12	Comp 4
Comp 5	-0,2	-0,3	0,2	-0,2	-0,1	-0,3	-0,3	-0,1	0,3	0,1	0,1	-0,2	0,2	0,2	0,2	-0,1	0,1	0,1	0,1	-0,2	-0,1	0,0	-0,2	-0,1	0,3	0,6	-2,7	-0,1	0,0	-0,3	12	Comp 5
Comp 6	0,3	0,2	0,0	0,1	0,0	0,2	-0,3	0,0	0,0	0,0	0,0	0,0	-0,1	-0,1	-0,1	-0,1	-0,1	-0,1	0,0	0,2	-0,2	0,0	0,1	0,0	0,2	-0,1	-1,1	-0,2	0,3	-0,7	7	Comp 6
Comp 7	-0,3	0,1	0,0	0,1	-0,1	-0,1	0,1	0,1	0,1	0,0	0,0	0,0	0,0	-0,1	0,0	0,5	0,0	0,0	0,0	0,2	0,3	-0,1	0,2	0,4	0,2	0,3	2,3	0,0	0,4	-0,1	8	Comp 7
Comp 8	-0,1	0,0	0,0	0,0	0,0	0,0	0,2	0,0	0,0	0,0	0,0	0,0	0,0	0,6	0,1	-0,1	0,0	0,0	0,0	0,1	-0,2	0,0	0,0	0,3	-0,1	0,0	4,5	-0,1	-0,4	-0,4	6	Comp 8
Comp 9	0,6	0,0	-0,1	0,0	0,0	-0,1	0,0	0,1	0,1	0,0	0,0	-0,1	0,0	0,2	0,0	-0,4	0,0	-0,1	0,0	0,1	0,1	0,1	0,0	0,2	0,2	0,2	2,2	0,0	0,2	0,3	6	Comp 9
Comp 10	-0,5	0,1	0,0	0,1	0,1	0,1	0,1	0,0	-0,2	0,0	0,0	0,1	-0,1	0,3	0,0	-0,4	0,0	0,0	0,0	-0,2	0,1	0,0	0,0	-0,1	0,4	-0,1	2,6	0,0	0,4	0,1	6	Comp 10
Comp 11	0,1	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,1	0,0	0,0	0,0	0,5	0,0	0,0	0,0	0,0	0,0	0,1	0,2	0,0	0,0	0,1	-0,4	-0,1	-4,9	0,1	0,3	0,0	4	Comp 11
Comp 12	-0,2	0,2	-0,2	0,1	0,0	0,1	0,0	0,2	0,1	0,0	0,0	0,1	-0,1	-0,2	-0,1	-0,3	-0,1	-0,1	-0,1	0,1	-0,1	0,0	0,1	0,3	0,0	0,3	-3,7	0,1	-0,3	0,0	6	Comp 12
Comp 13	0,1	0,1	0,0	0,0	0,2	0,1	-0,1	0,1	0,0	-0,1	-0,1	0,1	-0,1	0,2	-0,2	0,3	0,0	0,1	0,1	-0,3	-0,2	0,0	0,0	0,2	0,0	0,1	-1,0	-0,4	0,0	0,2	6	Comp 13
Comp 14	-0,1	0,0	0,0	0,1	-0,1	-0,1	0,2	0,0	0,1	0,0	0,0	-0,1	0,0	-0,1	0,1	-0,1	-0,1	-0,1	-0,1	0,1	-0,3	0,0	0,1	-0,3	-0,4	0,2	1,3	-0,4	0,2	0,1	7	Comp 14
Comp 15	-0,2	-0,1	0,0	0,0	0,0	-0,1	-0,3	0,0	0,2	0,2	0,0	0,0	0,0	-0,1	0,0	-0,2	0,0	0,1	0,1	0,1	0,1	0,1	-0,1	0,3	-0,1	-0,3	1,4	-0,2	0,1	0,1	6	Comp 15
Comp 16	0,0	0,0	-0,1	0,1	0,0	0,2	-0,2	0,0	-0,3	0,0	0,0	0,0	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,3	0,4	0,1	-0,2	-0,2	0,0	0,2	-9,6	-0,2	-0,1	0,0	8	Comp 16
Comp 17	0,0	0,1	0,2	-0,1	0,2	0,0	0,2	0,1	0,1	-0,1	0,0	0,1	0,0	0,0	-0,1	0,0	0,1	0,2	0,3	0,3	-0,1	-0,1	-0,1	-0,1	0,1	0,0	-6,5	0,0	0,0	0,0	6	Comp 17
Comp 18	0,0	-0,2	0,0	-0,1	0,0	0,0	-0,2	-0,1	-0,3	0,1	0,0	0,0	0,0	0,0	0,0	-0,1	0,1	0,1	0,2	0,2	-0,3	0,0	0,3	0,0	0,0	0,1	1,3	0,1	0,1	0,1	5	Comp 18
Comp 19	0,0	0,0	-0,1	0,1	0,0	0,0	-0,1	0,0	-0,1	-0,3	0,0	-0,1	0,1	0,0	0,0	0,0	0,0	0,0	-0,1	0,1	-0,2	-0,4	-0,3	0,1	0,0	0,0	1,5	0,1	0,1	0,0	4	Comp 19
Comp 20	-0,1	0,0	0,0	0,1	0,1	0,0	-0,1	0,0	0,1	-0,5	0,0	-0,1	0,1	0,0	0,1	0,0	0,1	0,0	0,0	0,0	0,0	0,3	0,1	0,0	0,0	-0,1	-2,9	0,0	0,0	0,0	3	Comp 20
Comp 21	0,0	-0,4	-0,2	0,0	0,1	0,3	0,1	-0,1	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	-1,4	0,0	0,1	0,0	4	Comp 21
Comp 22	0,0	0,1	-0,3	-0,1	-0,1	-0,2	0,1	-0,3	0,1	-0,1	-0,2	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,2	0,0	0,0	0,0	0,0	0,0	0,0	0,0	-1,3	0,0	0,0	-0,1	4	Comp 22
Comp 23	0,0	-0,2	0,2	0,1	-0,4	0,1	0,0	0,1	0,0	-0,1	-0,2	0,1	0,0	0,0	0,0	0,0	-0,1	-0,1	0,2	0,0	0,0	0,0	0,0	0,0	0,0	0,0	5,9	0,0	0,0	0,0	2	Comp 23
Comp 24	0,0	0,2	0,1	-0,2	-0,1	0,3	0,0	-0,2	0,1	0,0	-0,1	0,0	0,0	0,0	0,1	0,0	0,0	0,2	-0,2	0,0	-0,1	0,0	0,0	0,1	0,0	0,0	6,8	0,0	0,0	0,1	4	Comp 24
Comp 25	0,0	0,1	0,2	0,2	0,1	0,1	-0,1	-0,3	0,1	0,1	0,0	-0,2	0,0	0,0	0,1	0,0	-0,2	-0,2	0,2	0,0	0,0	0,0	0,0	0,0	0,0	0,0	1,0	0,0	0,0	0,1	3	Comp 25
Comp 26	-0,1	0,1	-0,1	-0,4	0,0	0,1	0,0	0,2	-0,1	-0,1	0,2	-0,1	0,0	0,0	0,1	0,0	-0,2	-0,2	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	2,5	0,0	0,0	0,0	3	Comp 26
Comp 27	0,0	0,0	-0,1	0,0	0,1	0,0	0,0	0,1	0,0	0,0	-0,2	-0,2	0,0	0,0	0,0	0,0	-0,3	0,3	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	8,6	0,0	0,0	0,0	5	Comp 27
Comp 28	0,0	0,0	0,0	0,0	0,1	0,0	0,0	0,1	0,0	0,1	-0,3	-0,2	0,0	0,0	0,0	0,0	0,3	-0,2	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	5,3	0,0	0,0	0,0	4	Comp 28
Comp 29	0,0	0,0	0,0	0,0	-0,1	0,0	0,0	-0,1	0,0	0,0	0,2	-0,2	-0,4	0,0	-0,2	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	-1,3	0,0	0,0	0,0	3	Comp 29
Comp 30	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,1	-0,3	0,0	0,4	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	-4,1	0,0	0,0	0,0	3	Comp 30

Legenda:

- Correlação Positiva superior a 20%
- Correlação entre -20% e 20%
- Correlação Negativa inferior a -20%