UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Tomada de decisões em sistemas financeiros utilizando algoritmos de aprendizado de máquina supervisionado

Luis Carlos Otte Junior

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

ICMC USP SÃO CARLOS

Luis Carlos Otte Junior

Tomada de decisões em sistemas financeiros utilizando algoritmos de aprendizado de máquina supervisionado

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

Área de Concentração: Matemática, Estatística e Computação

Orientador: Prof. Dr. André Carlos Ponce de Leon Ferreira de Carvalho

USP – São Carlos
Dezembro de 2018

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

O89t

Otte JUNIOR, Luis Carlos
Tomada de decisões em sistemas financeiros utilizando algoritmos de aprendizado de máquina supervisionado / Luis Carlos Otte JUNIOR; orientador André Carlos Ponce de Leon Ferreira de Carvalho. -- São Carlos, 2018.
75 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2018.

1. cobrança. 2. aprendizado máquina. 3. redes neurais. 4. árvores de decisão. 5. mineração de dados. I. Ponce de Leon Ferreira de Carvalho, André Carlos, orient. II. Título.

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

Luis Carlos Otte Junior

Decision making in financial systems using supervised machine learning algorithms

Master dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP, in partial fulfillment of the requirements for the degree of the Master – Professional Masters in Mathematics, Statistics and Computing Applied to Industry. *FINAL VERSION*

Concentration Area: Mathematics, Statistics and Computing

Advisor: Prof. Dr. André Carlos Ponce de Leon Ferreira de Carvalho

USP – São Carlos
December 2018

Este trabalho é dedicado minha esposa Maíra e meus filhos Lorena e Rodrigo
meus pais e cunhados que sempre me apoiaram e acreditaram em meus objetivos e sonhos.
Em especial, ao meu avô Paulo Botelho que me fez conhecer um computador quando criança.

AGRADECIMENTOS

Os agradecimentos principais são direcionados a André Ponce de Leon Carvalho e Rodrigo Mello que me auxiliaram a conclusão deste trabalho e ao coordenador Antônio Castelo que idealizou o curso MECAI.

"O dom do poder mental vem de Deus, do Ser Divino, e se concentrarmos nossas mentes nessa verdade, nos tornamos sintonizados com esse grande poder. Minha Mãe me ensinou a buscar toda a verdade na Bíblia."

(Nikola Tesla)

RESUMO

OTTE JUNIOR, L. C. **Tomada de decisões em sistemas financeiros utilizando algoritmos de aprendizado de máquina supervisionado**. 2018. 75 p. Dissertação (Mestrado – Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2018.

Embora existam soluções para sistemas de cobrança e telecomunicações que apresentem relatórios para auxílio à cobrança de clientes, ambas carecem de informações que apoiem a tomada de decisões, nas análises estratégicas e na propensão de pagamento. Desse modo, o objetivo deste projeto é implementar ferramentas e soluções inteligentes a fim de reduzir o desperdício de tempo e aumentar a produtividade do gestor, decorrentes da necessidade da análise e cruzamento de todos os dados para tomar qualquer ação durante os processos de cobrança e gestão de custos.

Palavras-chave: cobrança, aprendizado máquina, redes neurais, árvores de decisão, mineração de dados.

ABSTRACT

OTTE JUNIOR, L. C. **Decision making in financial systems using supervised machine learning algorithms**. 2018. 75 p. Dissertação (Mestrado – Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2018.

Although there are solutions for billing and telecommunications systems to present reports to support debt collection, both lack information to support decision making in strategic analysis and propensity to pay. Thus, the goal of this project is to implement intelligent tools and solutions taht are able to increase their productivity and reduce waste of managers time, due to the need of analyzing and crossing all the data to take action during the collection processes and cost management.

Keywords: credit recovery, machine learning, neural networks, decision trees, data mining.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ASR	Automatic speech recognition
CRM	Customer Relationship Management
MECAI	Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

Após o advento da Internet, a quantidade de informações geradas ano após ano tem crescido de maneira muito rápida a medida em que novos meios de comunicação estão ganhando destaques e automaticamente se popularizando. Ao mesmo tempo que essas informações ganham espaço no mercado, tanto na área de armazenamento quanto na área de processamento desses dados, que em grande parte não são estruturados, existe também uma parcela de dados estruturados presentes em sistemas de negócios e ferramentas analíticas. Essas ferramentas são utilizadas principalmente para descreverem comportamentos e padrões com o objetivo de mapear seus consumidores e facilitar tomadas de decisões.

Apesar do uso de técnicas básicas de inferência estatísticas em indicadores e relatórios, a complexidade e o grande volume de dados dificultam cada vez mais a perspectiva do funcionamento e eficiência dessas ferramentas (CABENA *et al.*, 1998), exigindo uma outra abordagem para descrever padrões em dados estruturados.

Atualmente, o processo de mineração de dados e as técnicas de aprendizado de máquina têm demonstrado um grande desempenho em encontrar padrões e ao mesmo tempo "aprendê-los", com o objetivo de descrever projeções futuras dos dados e classificá-los em categorias, que de forma probabilística pode auxiliar tomadas de decisões para uma determinada ação gerando eficiência nos mais diversos setores do mercado, que no caso desse trabalho especificamente o ramo de cobrança de dívidas financeiras (DASS, 2006) e (BHAMBRI, 2011).

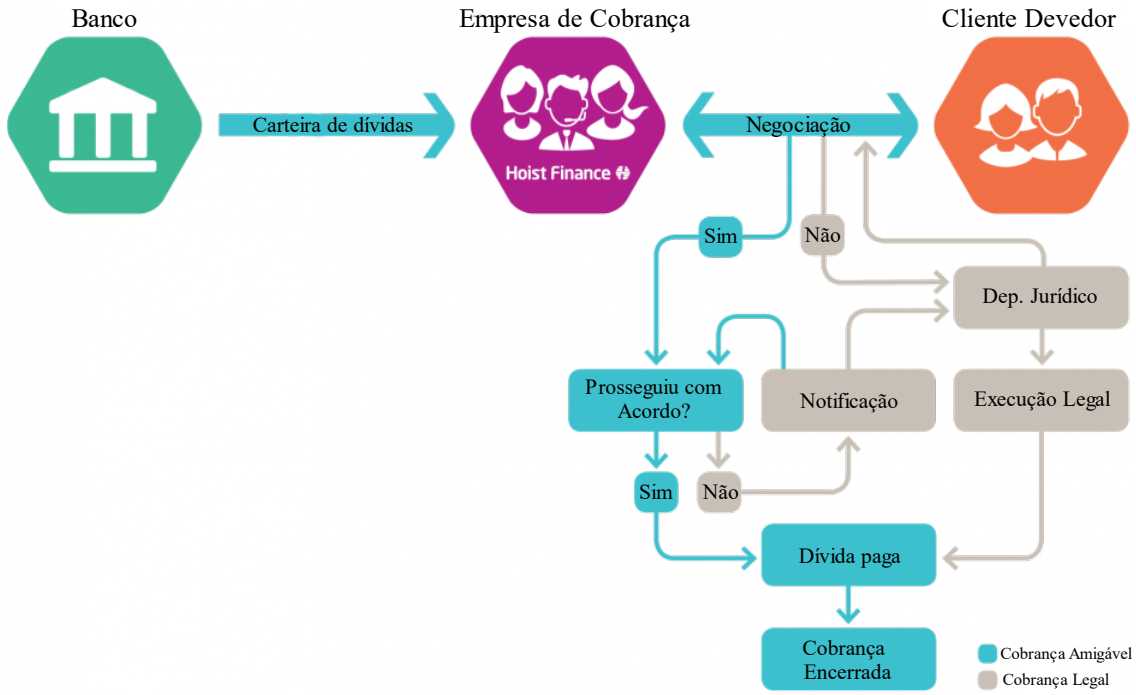
Uma grande parcela das dívidas que as pessoas incorrem não são reembolsadas em um tempo hábil. A definição de dívida é dada quando existe um valor em aberto a receber por um determinado credor que cedeu valores monetários por algum veículo de crédito ou modalidade, por exemplo cartões de crédito, financiamentos, empréstimos ou uma variedade de outros tipos de dívida ou obrigação de crédito. O débito é qualquer dívida que não foi reembolsada em sua data de vencimento, ou uma dívida em que um ou mais pagamentos parcelados não foram honrados. Os emissores de títulos normalmente empregam diferentes métodos para cobrar as

dívidas em atraso, no todo ou em parte.

Por exemplo, um cliente devedor que deixou de efetuar pagamentos mensais de sua fatura de cartão de crédito. Normalmente, a empresa de fornecimento de crédito utiliza diversos meios de cobrança, tais como cartas, SMS ou telefonemas (ZHANG; KISIELIUS, 2009), para incentivar o cliente a efetuar o pagamento.

No entanto, uma vez que o vencimento da conta atingiu 30 dias de atraso, dependendo do contrato da linha de crédito, o débito alcança a definição legal de uma dívida e deve ser liquidada. Para que essa dívida seja paga, são realizadas tarefas subsequentes de cobrança de dívidas que são conhecidas como "operação de cobrança". Neste momento, a empresa de fornecimento de crédito pode continuar a cobrar a dívida, ou pode optar por vender a dívida para uma agência de cobrança conforme processo ilustrado na Figura 1.

Figura 1 – Processo da cobrança



Fonte: Adaptada de Hoist Finance (2017).

Desse modo, as agências de cobrança ao receberem um portfólio de clientes devedores, elas possuem um prazo de dias determinado pelo banco para cobrarem estas dívidas e negociarem os pagamentos com os clientes devedores. Porém essas cartelas em geral possuem uma grande quantidade de clientes devedores dos quais a agência de cobrança é obrigada a realizar no mínimo um contato, porém os demais contatos mais conhecidos por acionamentos ficam a critério das estratégias de cobrança do gestor.

Em alguns casos o gestor pode até utilizar recursos de análise financeira aplicando réguas em *scores* ou mesmo realizar uma seleção manual dos clientes a serem trabalhados nas negociações e acionamentos. Mas essas ações não contam com dados históricos dos clientes

tais como movimentações financeiras e informações relevantes que o banco possa fornecer para facilitar esse processo, deixando a tomada de decisão de quais clientes os agentes cobradores devem cobrar apoiada em uma análise pessoal desse gestor.

O principal problema com relação à tomada de decisão apoiada pelo gestor é a baixa produtividade, que é possível observar nos dados da empresa envolvida neste trabalho conforme Tabela 1 os casos de sucesso *OK* e não sucesso *NOK* da agência de cobrança em um período do segundo semestre do ano de 2016 ao primeiro semestre de 2017.

Tabela 1 – Resumo das cobranças realizadas em 2016 e 2017 pela empresa A.

Semestre	Status	Qtde.	%	$\bar{x}$	$\widetilde{x}$	σ	ρ
2-2016	OK	3412	8	568,6	576,0	307,4	-0,6
1-2017	OK	5258	7	876,3	467,5	827,1	
2-2016	NOK	40893	92	6815,5	5423,5	4984,6	-0,7
1-2017	NOK	73096	93	12182,6	7571,5	12053,7	

Fonte: Dados da pesquisa.

Nota – Resumo em números dos casos de sucesso e não sucesso das cobranças realizadas por uma agência de cobrança, $\bar{x}$ é a média mensal, $\widetilde{x}$ é a mediana mensal, σ é o desvio padrão mensal e ρ é o índice de correlação de Pearson entre os totais de cada categoria para os dois semestres.

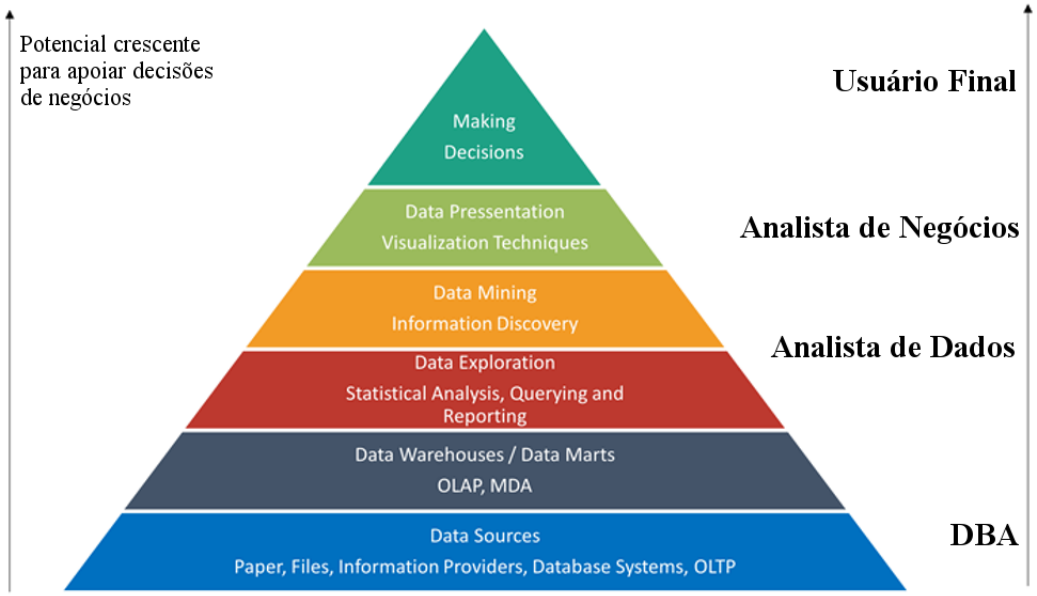
Apesar do grande volume de contratos, a empresa de cobrança realiza um esforço em cobrar todos de forma sistêmica e automatizada. Entretanto, isso geralmente ocorre sem uma maior eficiência, impossibilitando uma tomada de decisão precisa, que vise a redução dos custos das operações e aumento de suas receitas.

A baixa eficiência pode estar relacionada principalmente pela falta de uma análise mais consolidada e estratégica. Conforme ilustra o gráfico da Figura 2, grande parte do processo de cobrança realizado pela empresa, concentra-se apenas nos dois primeiros níveis, onde os dados estão dispostos em tabelas com milhões de registros. Esse processo demanda que um gestor de carteiras de cobrança crie estratégias de cobrança manualmente com base em seus milhares de dados, sem conseguir projetar estatisticamente suas cobranças futuras e muito menos visualizar padrões, devido a dimensionalidade e complexidade de informações neles presentes.

A partir dos demais níveis até o topo as informações passam a ser cada vez mais sumarizadas e representar melhor os grupos de clientes e demonstrar padrões que tornam possível uma tomada de decisão rápida e precisa.

Para isso, o emprego do processo de Mineração de Dados (FACELI *et al.*, 2011) aliado a uma abordagem de inteligência de negócios torna possível encontrar padrões nos dados

Figura 2 – Mineração de Dados e a Inteligência de Negócios



Fonte: Adaptada de Cabena *et al.* (1998).

estruturados. Isso permite reproduzir cenários futuros a partir de novas entrada de informações de cobranças o que, em contrapartida, garante uma tomada de decisão mais precisa e flexível no quesito de eficiência com um melhor apoio à tomada de decisões.

Com essa previsão, a empresa de cobrança consegue focar seus esforços de forma segmentada em suas carteiras mais lucrativas, com maior hipótese de serem negociadas com sucesso, e reduzir o custo nas carteiras menos lucrativas, com uma menor hipótese de sucesso.

1.1 Objetivo

O objetivo deste trabalho é o de projetar e desenvolver uma ferramenta de análise de dados utilizando algoritmos de aprendizado de máquina supervisionado, que indicará quais clientes devem ser priorizados por haver maior chance de sucesso na negociação, e consequente retorno de investimento (Wei Fan; Janek Mathuria; Chang-Tien Lu, 2005). Essa ferramenta deve ser integrada ao sistema de cobrança da empresa para estimar a propensão de pagamento de cada contrato em uma escala de 0 a 5, com o objetivo final de auxiliar a tomada de decisão na organização das cobranças a serem realizadas pelo *call center* da agência.

1.2 Contribuição do trabalho

A principal contribuição deste trabalho é de demonstrar a possibilidade de utilizar os dados de cobranças realizadas pelo cliente para induzir modelos preditivos a partir de algoritmos de aprendizado de máquina.

Com uma abordagem nas áreas de mineração de dados, serão tratados casos de ruídos, valores ausentes e classes desbalanceadas demonstrando o processo de preparo dos dados. Posteriormente a essa etapa, com alguns ajustes nos parâmetros das técnicas, os algoritmos deverão ser capazes de encontrar padrões em casos de sucesso e não sucesso.

O trabalho foi desenvolvido de forma a ser aplicado ao ambiente da empresa de cobrança como uma ferramenta para cálculo de escores de acordo com a hipótese da cobrança ser efetivada com sucesso. A ferramenta deve ser integrada como um módulo ao software *Customer Relationship Management* (CRM) do cliente. Desta forma, a tomada de decisão do gestor de cobrança deve tornar mais eficiente o emprego de uma abordagem de automação de discagem e cobranças por agentes.

1.3 Descrição dos capítulos

Este trabalho está estruturado com seis capítulos organizados seguindo os passos e conceitos da metodologia de Mineração de Dados.

As partes estão organizadas da seguinte forma:

No Capítulo 2 analisa-se a literatura relacionada, tanto nos campos do financiamento de crédito e cobrança quanto em análise de negócios. Apresenta-se também os trabalhos anteriores sobre análise e gerenciamento de contas e dívidas em atraso, bem como a forma em que as técnicas de aprendizado da máquina foram aplicadas em campos similares de recuperação de dívidas.

No Capítulo 3 apresenta-se como os dados foram processados nesse trabalho bem como análises estatísticas realizadas com o objetivo de fornecer uma visão geral dos dados.

Após essas primeiras abordagens com os dados, foi possível construir e ajustar um modelo preditivo capaz de calcular scores de eficiência de uma futura cobrança de cada dívida no Capítulo 4 e no Capítulo 5, comparar o desempenho de vários algoritmos de aprendizado de máquina aplicados ao problema em particular da previsão do *score* de cobrança.

No Capítulo 6 por fim, mostra-se a conclusão e também dispõe de algumas ideias de trabalho adicional.

CAPÍTULO

2

CONCEITOS E TRABALHOS RELACIONADOS

Neste capítulo apresenta-se alguns estudos relacionados à aplicação de técnicas de aprendizado de máquina no setor financeiro, que servirão de base para a compreensão da área a ser explorada pela pesquisa. O principal objetivo é encontrar um padrão na forma de utilização das técnicas como também descobrir qual técnica é mais empregada e pode ser utilizada no trabalho.

2.1 Principais trabalhos relacionados

Ao realizar a revisão bibliográfica, foram encontrados diversos trabalhos relacionados a área de inteligência artificial aplicada a finanças. Desses, foram selecionados os trabalhos que possuíam maior relevância para a área de cobrança e recuperação de créditos. Outra questão importante considerada como critério de seleção dos trabalhos foi considerar que os dados utilizados são de origem de sistemas de gestão de relacionamento com clientes CRM, pois, essa questão também está ligada a proposta de trabalhar com dados já estruturados originados de um CRM.

Os modelos de *Credit Score* são os mais comuns e utilizados atualmente (KUMAR *et al.*, 2005). Em geral, são disponibilizados através de serviços de consulta fornecidos por empresas privadas como, SERASA, SPC e Equifax do Brasil (Castelar Pinheiro; MOURA, 2003), ou implementados com dados sociodemográficos dos clientes, como idade, sexo, renda entre outras que são solicitadas no momento da aquisição do crédito. Por meio do modelo de *credit score* é possível prever a probabilidade de um novo cliente atrasar o pagamento da linha de crédito por N dias e projetá-la nos meses seguintes (ANDERSON, 2007) ou até mesmo obter uma variável resposta bom ou mau cliente (DERELIOUGLU; GÜRGEN; OKAY, 2009)

Estes trabalhos (DERELIOUGLU; GÜRGEN; OKAY, 2009), (SZCZERBA; CIEMSKI,

2009) e (PEDRO; PROSERPIO; OLIVER, 2015) utilizaram técnicas de redes neurais, árvores de decisões e *Gradient Boosted Trees*. Os dados utilizados para pesquisa foram históricos relacionados a movimentações financeiras, dados sociodemográficos e registro das ligações telefônicas dos clientes pertencentes a uma base de dados do CRM utilizado pela empresa.

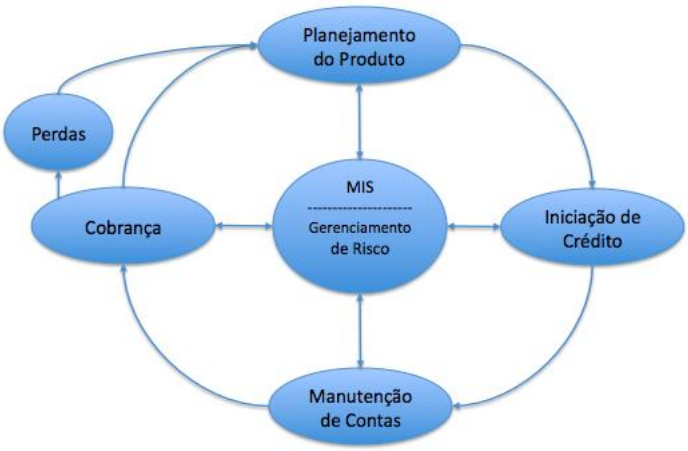
Outro modelo semelhante e considerado como uma variação do *Credit Score* é o *Behavior Score* ou Escore de Comportamento. Além do uso de dados sociodemográficos o modelo utiliza também dados comportamentais originados de históricos dos clientes presentes no portfólio da empresa como, histórico de pagamentos em dia, em atraso, quantidade de empréstimos, entre outros.

Ambos os modelos abordam o cliente como um ponto principal de análise, no caso do *Behavior Score*, a variável resposta é representada pelo risco do cliente se tornar um devedor. Esse modelo é considerado mais robusto que o modelo de *Credit Score* pois utiliza dados observados dentro do portfólio da empresa e não somente por alguns atributos fornecidos por departamentos de análise de crédito como o Serasa Experian ou apenas dados sociodemográficos. O trabalho (FERREIRA *et al.*, 2013), utilizou Redes Neurais para desenvolvimento do modelo e dados sociodemográficos e comportamentais dos clientes de um banco.

2.2 Ciclo de crédito ao consumidor

O uso de modelos relacionados a questão da concessão do crédito não se limita somente ao momento inicial a ser avaliado conforme os trabalhos referenciados anteriormente, mas em todo o processo posterior do fornecimento de crédito, ou seja, em quase todas as etapas do ciclo de crédito ao consumidor (LAWRENCE, 1984), ilustrado na Figura 3 é possível aplicar modelos com o objetivo de gerar eficiência e reduzir custos.

Figura 3 – O Ciclo de Crédito ao Consumidor



Fonte: Adaptada de LAWRENCE (1984).

Sendo assim, os modelos *credit score* são aplicados na segunda etapa do ciclo, que

consiste na venda do produto de crédito ao cliente.

A terceira etapa, a de Manutenção de Contas é considerada por (LAWRENCE, 1984) a mais complexa, pois é encarregada de manter os clientes satisfeitos, sempre lucrativos em seus pagamentos por parte da credora e evitar seus desligamentos ou cancelamentos. Nessa etapa são utilizados, por exemplo, modelos de previsão de cancelamento (*Churn*) e de retenção de clientes.

O modelo de previsão de cancelamento (*Churn*) é definido pela propensão de um determinado cliente deixar de realizar negócios com a empresa após um dado período.

Em geral, existe um grau dificuldade maior em modelar esse tipo de problema, devido ao alto nível de ruídos nos dados e problemas de desbalanceamentos das classes, pois em todos os casos, o número de cancelamentos será sempre mínimo e inferior. Os trabalhos (XIE *et al.*, 2009) e (FARQUAD; RAVI; RAJU, 2014) desenvolveram soluções para a questão do desbalanceamento.

O primeiro trabalho implementou um algoritmo híbrido baseado na técnica Random-Forest utilizando pesos nas classes menores e posteriormente aplicou um balanceamento de *oversampling*. A técnica que tem como objetivo replicar as classes minoritárias forçando um balanceamento das classes.

No segundo trabalho, foi utilizada a técnica de SVM + Naive-Bayes Tree para extração de regras e classificação. Os dados foram balanceados por meio da técnica de SMOTE (CHAWLA *et al.*, 2002).

Já o modelo de retenção de clientes, embora pareça semelhante ao modelo de *Churn*, difere na proposta de quantificar se um determinado cliente está propenso ao processo de cancelamento ou uma nova aquisição de produto *"next buy"* (LARIVIERE; POEL, 2005). Esse trabalho comparou os algoritmos de regressão logística e RandomForest para classificação das classes 'Próxima Compra', 'Em Processo de Cancelamento' e 'Queda de Lucro' tratando-se de uma abordagem multi-classes.

A quarta etapa seguinte do ciclo de crédito é a cobrança, segundo (LAWRENCE, 1984), mesmo que todo esforço seja realizado para concessões de créditos saudáveis sem alto risco é inevitável que alguns clientes não cumpram suas obrigações de pagamento dentro dos prazos de vencimentos. Os clientes podem tornar-se inadimplentes por diversos motivos, desde o mais simples como esquecimento da data de vencimento, até situações mais delicadas como perda da fonte de renda ou endividamento por falta de controle orçamentário resultando em uma insolvência.

2.3 Aplicações de Aprendizado de Máquina

Com relação a modelos de aprendizado de máquina é possível encontrar inúmeros tipos e aplicações, onde de forma sistêmica estão sempre relacionados aos processos básicos de cobrança,

que são divididos em faixas de dias em atraso conforme políticas de crédito particulares de cada empresa (Silva Santo, 2013).

O primeiro processo é caracterizado por clientes que não pagaram seus compromissos após prazo de vencimento com um atraso inferior à 180 dias. Os modelos de *Late Payment* desempenham um papel fundamental para solucionar essa questão. Sua característica é prever o quão tarde será o pagamento da dívida permitindo uma cobrança amigável, se a dívida ainda está dentro da empresa no departamento de contas a receber. A técnica de aprendizado de máquina utilizada nesse trabalho (ZENG *et al.*, 2008) foi a de árvores de decisão C4.5.

O segundo processo é iniciado após o atraso exceder os 180 dias, em geral a empresa envia uma notificação formal informando a negativação nos serviços de proteção ao crédito, porém não apenas faz isso como também continua a cobrar o cliente. Em alguns casos a dívida é transferida para uma empresa especializada em cobranças (MEDEIROS; BRITO; ARAUJO, 2008), sendo que nessa etapa, pode ser aplicado um modelo chamado *Collection Score*, o qual será detalhado mais adiante no contexto dessa etapa.

Uma última etapa auxiliar vinculada a cobrança chamada de perda, está relacionada a dívidas que não foram recuperadas com um prazo superior à 180 dias, são dadas como perdas, porém ficam a espera de uma posterior ação de recuperação de crédito, em negociações de descontos de juros e parcelamento da dívida.

Uma etapa localizada no centro do ciclo chamada Sistema de Informações Gerenciais ou (MIS, do inglês Management Information System). Ela é responsável por gerenciar as transições e operações entre as etapas.

O MIS tem um papel fundamental na definição das estratégias de cada etapa com base nos modelos geridos por ela, sendo uma espécie de *hub* central permitindo a troca de dados de todas as etapas com os gestores, vendedores e cobradores.

Ao concluir a análise do ciclo de crédito, foi possível observar que o trabalho que melhor se aproxima com a questão do problema a ser resolvido pelo trabalho é o modelo de *Collection Score*. O modelo busca identificar possíveis clientes pagadores de suas dívidas já em um cenário em que todos são devedores, diferente dos modelos de *Credit* ou *Behaviour Score* onde o cliente pode ser ou não devedor.

Para utilizar o modelo de *Collection Score* são considerados dados associados ao relacionamento com a empresa, por exemplo histórico de cobranças, custos envolvidos no processo de cobrança, dados sociodemográficos, comportamentais entre outros atributos presentes no banco de dados do CRM da empresa de cobrança e não somente no banco credor. Na maioria dos casos, a ação de cobrança é realizada de forma externa e terceirizada, o banco credor cede sua cartela de clientes devedores para agências especialistas em recuperação de crédito e negociações (MEDEIROS; BRITO; ARAUJO, 2008) de portfólios de créditos não-performados ou conhecidos como (NPLS, do inglês Non-performing loans).

Os trabalhos (MACHADO, 2016) e (GONÇALVES; GOUVÊA, 2015) utilizaram uma abordagem clássica empregando a técnica de regressão logística para classificação de bons e maus pagadores e trabalharam com a hipótese de que a representação das classes dos dados estarão balanceadas.

O primeiro trabalho (MACHADO, 2016), apresentou as classes desbalanceadas e não aplicou uma técnica para balanceamento o que possivelmente influenciou no resultado do modelo demonstrado que a técnica de regressão logística só conseguiu representar a classe majoritária. A consequência disso foi uma acurácia muito baixa para a classe minoritária.

Um ponto importante a considerar é que nos trabalhos anteriores das outras etapas do ciclo, sempre existiram casos de classes minoritárias (maus pagadores e cancelamentos) expressando um cenário real do mercado e em alguns casos ocorreram etapas para o balanceamento das classes.

O trabalho (GRIGORCHUK *et al.*, 2015) utilizou a técnica Tobit Type II e apresentou uma boa acurácia, suas classes não estavam tão desbalanceadas quanto os trabalhos relacionados a área de cobrança.

Para melhor compreensão de todos os modelos foi elaborada a Tabela 2, com o objetivo de comparar as escolhas de cada trabalho com os respectivos resultados.

Tabela 2 – Resultado da acurácia dos classificadores nos trabalhos pesquisados

	Trabalhos										
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]
Modelos	Credit			Behavior	Churn		Retenção	Late Payment	Collection		
Regr. Logística			0,72	0,87			0,69	0,70	0,78	0,83	
Árvore Decisão		0,09*		0,85	0,62			0,79			
MLP	0,75			0,89	0,78						
SVM	0,75		0,72		0,87	0,86					
k-NN	0,74										
GB Tree			0,72					0,67			
RandomForest					0,93		0,73				
NBTree						0,94					
SVM+NBTree						0,83					
Tobit Type II											0,80
Dist. Classes %	72 / 28	91 / 9	80 / 20	91 / 9	95 / 5	93 / 7	84 / 16	65 / 35	82 / 18	50 / 50	60 / 40

Fonte: Dados da pesquisa.

Nota – [1] (DERELIOUGLU; GÜRGEN; OKAY, 2009), [2] (SZCZERBA; CIEMSKI, 2009), [3] (PEDRO; PROSERPIO; OLIVER, 2015), [4] (FERREIRA *et al.*, 2013), [5] (XIE *et al.*, 2009), [6] (FARQUAD; RAVI; RAJU, 2014), [7] (LARIVIERE; POEL, 2005), [8] (ZENG *et al.*, 2008), [9] (MACHADO, 2016), [10] (GONÇALVES; GOUVÊA, 2015), [11] (GRIGORCHUK *et al.*, 2015).

Ao observar a Tabela 2 é possível verificar que 80% dos trabalhos utilizaram dados com

a distribuição das classes desbalanceadas principalmente os trabalhos relacionados aos modelos *Churn*, *Behavior Score* e *Collection Score*.

No trabalho (FARQUAD; RAVI; RAJU, 2014), a utilização da técnica de SMOTE apresentou melhor resultado com relação à técnica de *oversample*. Já o trabalho (XIE *et al.*, 2009) desenvolveu um balanceamento utilizando pesos maiores nas classes minoritárias e conseguiu resultados superiores por meio do algoritmo de aprendizado de máquina RandomForest. Os trabalhos (FERREIRA *et al.*, 2013) e (MACHADO, 2016) não aplicaram nenhuma técnica de balanceamento de classes consequentemente não apresentando resultados semelhantes aos anteriores pesquisados.

Conforme a proposta do trabalho de implementar um modelo de *Collection Score* com dados desbalanceados, foi possível concluir ao pesquisar alguns trabalhos relacionados diretamente e indiretamente a área, de que é necessário um balanceamento das classes.

O trabalho de (MACHADO, 2016) não realizou tal procedimento que acabou influenciando o modelo de regressão logística de forma negativa a ajustar-se somente nas classes negativas. Conforme matriz de confusão a precisão dos verdadeiros positivos foi de somente 33%, ou seja, o modelo não se ajustou corretamente as classes positivas.

Desse modo, foi possível eleger algumas técnicas de Aprendizado de Máquina para aplicação no trabalho. Serão utilizadas na pesquisa as seguintes técnicas: Random Forest, Árvores de Decisão e Redes Neurais Perceptron Multicamadas. A técnica Regressão Logística também será utilizada devido sua presença em quase todos os trabalhos pesquisados.

CAPÍTULO

3

DADOS E PRÉ-PROCESSAMENTO

Será abordado nesse capítulo, algumas análises e operações de pré-processamento dos dados coletados diretamente do banco de dados da agência de cobrança.

Embora existam algoritmos de aprendizado de máquina que desempenhem um papel rápido e simples com relação à função de extração de conhecimento do conjunto de dados. O desempenho, em geral, pode ser afetado pelo estado dos dados se não existir uma etapa dedicada ao pré-processamento e análise do conjunto de dados (FACELI *et al.*, 2011).

O pré-processamento dos dados é uma etapa fundamental e influência todo o processo de aprendizado de máquina que age de forma pragmática e empírica. Se os dados não contemplam as informações necessárias para reproduzir um melhor espaço solução, os algoritmos tendem a não se ajustarem aos dados, efeito chamado *underfitting*, ou os modelos tendem a "decorar"as informações, efeito chamado *overfitting*.

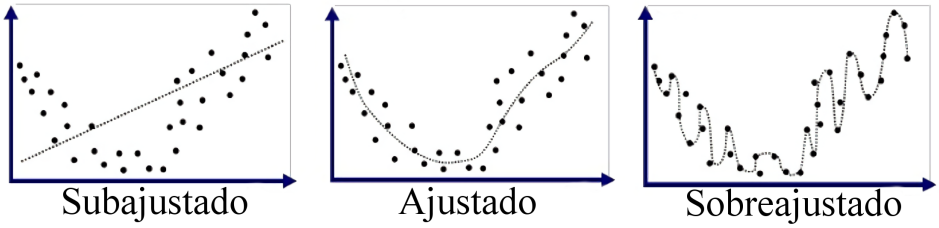
Ambos os efeitos geralmente podem ser resultados de um insucesso no processo de modelagem dos dados ou o não estabelecimento de um limite no grau de liberdade do algoritmo, onde quanto maior esse grau, maior será a variança e menor será o viés.

Ao deixar um algoritmo de árvore de decisão que possui uma estratégia *greedy* sem limites por exemplo, pode ocasionar um *overfitting* conforme ilustração da Figura 4. Para evitar esse tipo de problema é necessário encontrar um *equilíbrio no viés-variância* (GEMAN; BIENENSTOCK; DOURSAT, 1992) conhecendo o conjunto de dados e trabalhando no processo de refinamento e melhoria.

Na etapa de pré-processamento, os primeiros esforços são voltados a compreender os dados. Eles podem apresentar diferentes características, dimensões e formatos, onde os atributos discretos ou contínuos, ou seja, valores do tipo simbólicos, texto ou numéricos, por exemplo, um número real.

Além do tipo dos dados, outra questão importante a ser considerada no processo é a

Figura 4 – *Underfitting* (viés alto, baixa variância) vs *Overfitting* (viés baixo, alta variância)



Fonte: Adaptada de Liew (2016).

validação e análise da integridade dos dados, pois, em grande parte dos casos podem conter ruídos e imperfeições, com valores incorretos, inconsistentes, duplicados ou ausentes.

Por último antes de aplicar os dados a um modelo é necessário em alguns casos, utilizar técnicas de amostragem, tratamentos para dados desbalanceados, modificações para adequação dos tipos de atributo, limpeza dos dados, integração de dados, transformações dos dados e redução de dimensionalidade (FACELI *et al.*, 2011).

Algumas abordagens citadas anteriormente serão aplicadas de modo a transformar os dados em um conjunto que expresse uma boa representatividade do problema em questão e redução de custos em termos de processamento.

Portanto, para obter uma boa extração de conhecimento do conjunto de dados é preciso inicialmente realizar uma análise e refinamento assim amenizando possíveis erros de generalização dos algoritmos de classificação, pois, para a solução de um problema utilizando técnicas de aprendizado de máquina é necessário formar um conjunto de dados que melhor representam o espaço solução.

3.1 Extração dos dados

Os dados foram extraídos diretamente do banco de dados da empresa de cobrança que possui mais de 100 tabelas, porém, apenas 5 tabelas foram necessárias para reunir dados pertencentes a duas categorias de acordo com a necessidade do trabalho.

A primeira categoria é a de dados relacionados ao processo de dívida do qual a empresa recebe dados de contratos e passa a trabalhar com esses dados durante as cobranças, esses dados são fornecidos diretamente pelo banco responsável pela carteira de cobrança.

A segunda categoria está associada aos dados dos clientes devedores vinculados aos contratos, com a natureza de pessoa física ou jurídica.

Outros dados também estão disponíveis, por exemplo, o histórico do processo de cobrança, onde contém dados dos cobradores e as formas que a cobrança foi realizada, mas esses dados foram descartados, pois o objetivo principal é construir um conjunto de dados para o

aprendizado de máquina *batch* ou *off-line* (BOTTOU, 1998) (SHALEV-SHWARTZ *et al.*, 2012). Essa abordagem foi escolhida devido o fato que de esses dados não estarão presentes no momento inicial do recebimento da carteira de cobrança e o modelo a ser desenvolvido pelo trabalho irá classificar previamente os contratos com o *Collection Score*.

Uma vez que os dados são extraídos e disponibilizados de forma estrutural, mais especificamente em uma tabela melhor representado por uma matriz de dados $\mathbf{X}_{n\times d}$, as colunas são denominadas como atributos e as linhas como objetos.

O período escolhido para extração foi de junho de 2016 à junho de 2017, contemplando as épocas de sazonalidades da economia, resultando em um total de 141528 objetos e 190 atributos.

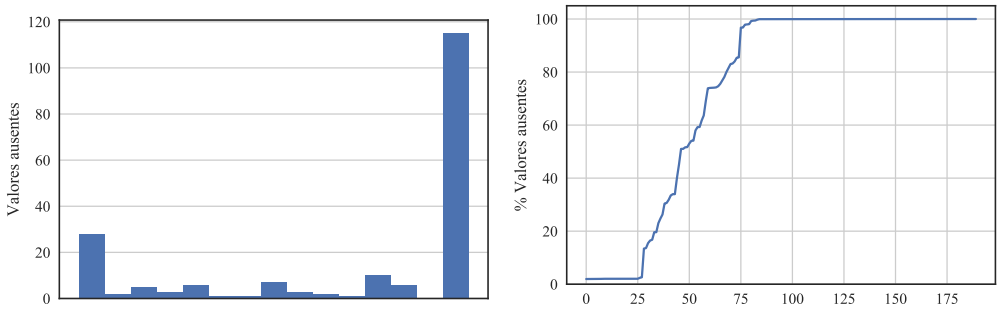
3.2 Seleção dos atributos

Em grande parte das aplicações reais de classificação, aproximação, previsão e otimização as bases de dados originadas de SGDB ou data warehouse (KENNETH; LAUDON, 2007) contêm um amplo número de atributos. Todavia, geralmente alguns atributos não possuem relevância significativa e podem até mesmo ser redundantes. Logo, um problema comum em aplicações reais é a seleção dos atributos, porém, um desafio ainda maior é deparar-se com atributos com nomes genéricos ou de comportamento abstrato. Essas situações dificultam o processo de seleção que se realizado de forma manual sem auxílio de nenhuma técnica, pode resultar em perda de informações importantes.

O presente conjunto de dados contém 190 atributos, que podem estar sub-utilizados com poucas informações, vazios, existentes com o mesmo nome em outras tabelas ou até mesmo com nomes diferentes, porém, com conteúdo duplicado.

Com o objetivo de reduzir o número de atributos, a primeira análise realizada no conjunto de dados foi a de quantidade de valores ausentes por atributos ou colunas, conforme Figura 5.

Figura 5 – Histograma e distribuição dos valores ausentes por atributos



Fonte: Elaborada pelo autor.

Desse modo, é possível concluir que existem atributos que estão com mais de 60% de

valores ausentes (STRAUSS; ATANASSOV; De Oliveira, 2003), desse modo esses atributos foram removidos resultando em uma redução de 72% removendo 145 atributos.

Após a remoção dos atributos com alto número de valores ausentes restaram 55 dos 190, demostrando o quanto o modelo seria influenciado de forma negativa com muitos atributos que não representam o espaço solução (CLAVEL; MERCERON; ESCARGUEL, 2014).

Além dos valores ausentes, outra questão importante analisada foi a duplicidade de atributos, que inicialmente foi possível encontrar pelos nomes repetidos presentes nas tabelas selecionadas no processo de extração resultando em um total de 14 atributos repetidos que também foram removidos.

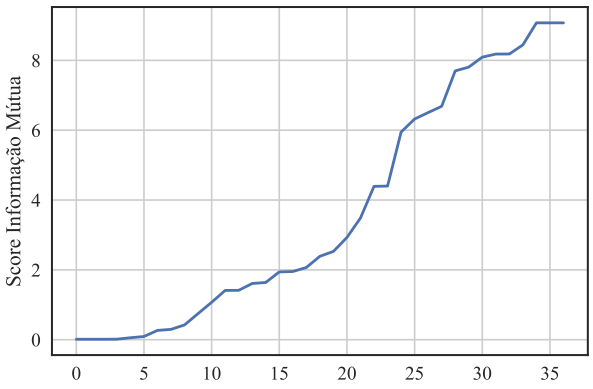
Após a análise de duplicidade por nomes, foram encontrados atributos com nomes diferentes, mas com conteúdo semelhante, sendo assim foi necessário utilizar uma técnica para avaliar a semelhança do conteúdo de todos os atributos com o objetivo de quantificar essa semelhança e realizar uma seleção.

A técnica adotada foi baseada no conceito de seleção de atributos através da análise do conteúdo dos atributos por meio do cálculo de Informação Mútua (BATTITI, 1994).

Para realizar a seleção foi calculada o valor da Informação Mútua entre cada coluna, gerando uma matriz de 41x41 representada conforme a Figura 8.

Ao observar a Figura 8, foi possível concluir que algumas colunas estão com valores de Informação Mútua altos com relação aos demais atributos, para validar essa possibilidade de forma visual, foi construído um gráfico com a distribuição valores calculados conforme a Figura 6.

Figura 6 – Distribuição dos valores de Informação Mútua dos atributos.

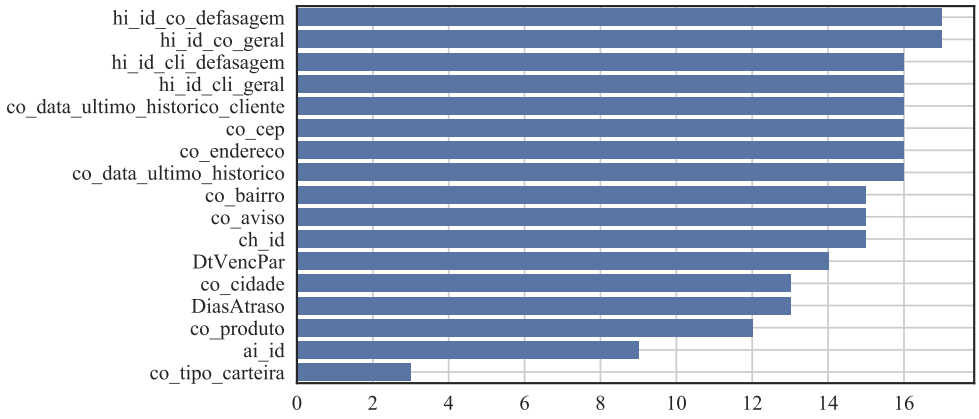


Fonte: Elaborada pelo autor.

Ao observar a Figura 6, foi possível verificar uma mudança abrupta nos valores de Informação Mútua a partir do valor 4. Definindo esse valor como critério para realizar a seleção dos atributos, foram selecionados todos valores acima de 4 e contabilizadas suas repetições de modo a encontrar os atributos que possuem maior semelhança entre todos os presentes.

Desse modo, foi possível encontrar e remover dois atributos com aspectos muito semelhantes, conforme Figura 7 com relação aos demais.

Figura 7 – Atributos com maior valor de Informação Mútua e maior número de repetições entre todos os atributos.



Fonte: Elaborada pelo autor.

Após as análises de valores ausentes e redundância dos atributos restaram 39 de 190 presentes antes a avaliação. Uma redução significativa em termos computacionais para o processamento dos algorítimos de classificação como também uma melhor representação de generalização do problema.

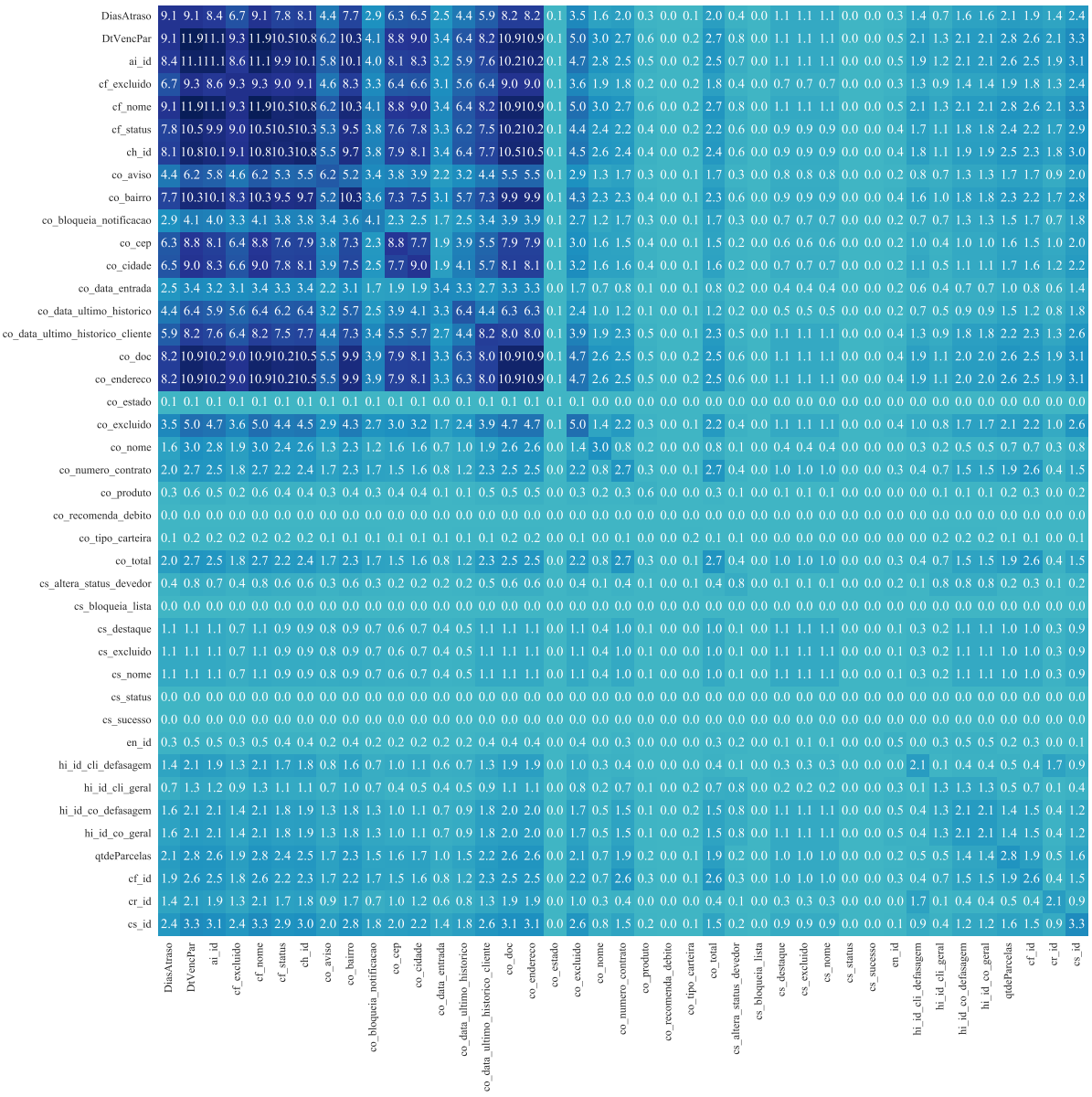
A segmentação das informações em altas dimensões tendem a desfavorecer alguns algoritmos no processo de convergência da solução como também reproduzir um modelo não ajustado (LIANG; TSAI; WU, 2015), problema tratado na área de mineração de dados como a "maldição da dimensionalidade". Nesses casos, mais pode ser menos, isto é, se você acrescentar dados demais você perde capacidade discriminativa e de avaliação (GAMA *et al.*, 2015).

3.3 Conhecendo os dados

Conforme citado anteriormente, um conjunto de dados é formado por atributos e objetos, sendo que cada objeto possui propriedades ou características que os descrevem por meio dos atributos, chamados também de atributos preditivos ou de entrada e que são constituídos por dois tipos gerais de domínio de dados: Qualitativo ou Nominal e Quantitativo ou Numérico (HANCOCK, 2011).

O primeiro tipo está associado a informações que "rotulam"ou "nomeiam"o objeto, por exemplo, um dos atributos do conjunto de dados *TipoPessoa* que descreve se o objeto é uma pessoa física "F"ou jurídica "J". Outro exemplo de atributos são nomes, endereços, categorias, tipos de dívidas, CEP e CPF. Existem outros que possuem uma função de qualificar o objeto e não são compostos de dados textos, por exemplo, o CEP que é uma informação numérica.

Figura 8 – Matriz dos valores de Informação Mútua calculada entre cada coluna.



Fonte: Elaborada pelo autor.

Os atributos numéricos contêm informações que são expressadas por números, ou seja, valor da dívida, número de parcelas entre outros que representam dados que podem ser aplicados em operações aritméticas. Eles diferem dos nominais, pois não é possível somar dois CPF ou subtrair CEP. Dentro dessa categoria existem duas formas ou escalas de representação dos dados, sendo a primeira Intervalar ou Discreto e Contínua ou Racional (*Ratio*).

Os dados intervalares são representados por datas e quantidade de parcelas, pois, possuem formas de representações de "distâncias"entre os números e somente algumas operações aritméticas de soma e subtração podem ser aplicadas por se tratar de uma representação de números inteiros. Já os dados contínuos estão vinculados, por exemplo ao valor da dívida entre outras informações que permitem aplicar todas as operações aritméticas.

A Tabela 3 descreve os tipos de dados de cada atributo presente no conjunto de dados, esse tipo de qualificação é de extrema importância para os passos seguintes até as aplicações as técnicas de aprendizado de máquina. Algumas delas não trabalham com dados nominais sendo necessária uma etapa de conversão desses dados (HANCOCK, 2011).

Outro atributo importante a destacar é o atributo alvo, que pode ser representado por um dos tipos de dados nominal ou numérico e só está presente em um conjunto de treino. Do mesmo modo anteriormente citado o tipo de atributo alvo pode influenciar na escola da técnica ou torna-se necessário fazer uma conversão.

Ao analisar de forma geral os atributos e qualificar seus respectivos tipos, foi possível concluir que alguns deles não são necessários e não expressarão o problema estatisticamente ou ainda estão em duplicidade, por exemplo os atributos *cf_id* e *cf_nome* são referências entre si, ou os atributos *ai_id*, *co_doc*, *co_id*, *co_nome* estão associados com identificadores únicos do objeto.

Um aspecto importante observado na análise foi a presença de outros possíveis atributos alvos, uma vez que o atributo definido pelo gestor da carteira como um possível sucesso na cobrança ou não sucesso é o atributo *cs_sucesso* que é binário, porém existe ainda os atributos *cs_id*, *cs_destaque* e *cs_nome* que representam de forma detalhada esse atributo alvo e foram removidos pois não existem em novos casos e podem influenciar negativamente o aprendizado de máquina.

Outro atributo encontrado que foi removido é o *co_aviso* onde possuí observações e relatos em forma de textos não estruturados e exigem um processamento de linguagem natural para extração correta de informações relevantes.

Atributos com valores únicos, como *cf_excluido* e *cf_status*, também foram removidos, pois não expressam conteúdos relevantes para explicar comportamentos.

Tabela 3 – Tipos e escalas de dados dos atributos presentes no conjunto de dados da pesquisa.

Atributo	Domínio	Escala
DiasAtraso	Quantitativo Discreto	Racional
DtVencPar	Quantitativo Discreto	Intervalar
ai_id	Qualitativo	Nominal
cf_excluido	Qualitativo	Nominal
cf_id	Qualitativo	Nominal
cf_nome	Qualitativo	Nominal
cf_status	Qualitativo	Nominal
ch_id	Qualitativo	Nominal
co_aviso	Qualitativo	Nominal
co_bairro	Qualitativo	Nominal
co_bloqueia_notificacao	Qualitativo	Nominal
co_cep	Qualitativo	Nominal
co_cidade	Qualitativo	Nominal
co_data_entrada	Quantitativo Discreto	Intervalar
co_data_ultimo_historico	Quantitativo Discreto	Intervalar
co_data_ultimo_historico_cliente	Quantitativo Discreto	Intervalar
co_doc	Qualitativo	Nominal
co_endereco	Qualitativo	Nominal
co_estado	Qualitativo	Nominal
co_excluido	Qualitativo	Nominal
co_nome	Qualitativo	Nominal
co_numero_contrato	Qualitativo	Nominal
co_produto	Qualitativo	Nominal
co_recomenda_debito	Qualitativo	Nominal
co_total	Quantitativo contínuo	Racional
cs_altera_status_devedor	Qualitativo	Nominal
cs_bloqueia_lista	Qualitativo	Nominal
cs_destaque	Qualitativo	Nominal
cs_excluido	Qualitativo	Nominal
cs_nome	Qualitativo	Nominal
cs_sucesso	Qualitativo	Nominal
en_id	Qualitativo	Nominal
hi_id_cli_defasagem	Qualitativo	Nominal
hi_id_cli_geral	Qualitativo	Nominal
hi_id_co_defasagem	Qualitativo	Nominal
hi_id_co_geral	Qualitativo	Nominal
qtdeParcelas	Quantitativo discreto	Intervalar
cf_id	Qualitativo	Nominal
cr_id	Qualitativo	Nominal
cs_id	Qualitativo	Nominal

Fonte: Dados da pesquisa.

Os atributos relacionados ao endereço do contrato *co_bairro*, *co_endereco*, *co_cep* e dados específicos de cada contrato *co_numero_contrato* e informações de acionamentos *co_data_ultimo_historico*, *hi_id_cli_defasagem*, *hi_id_cli_geral*, *co_produto*, *co_data_ultimo_-*

historico_cliente, *ch_id* também foram removidos.

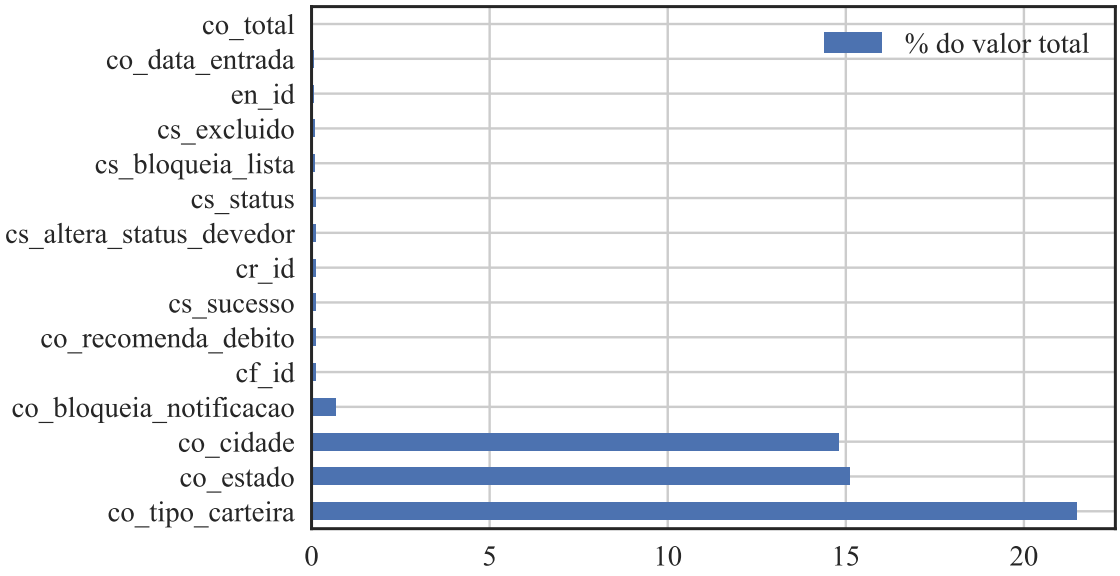
Por último, os atributos relacionados a data, como *DtVencPar* e *co_data_entrada*, serão mantidos, porém terão extraídos o ano e mês de referência e mantido o dia, para não expressar uma ocorrência ou correlação com o tempo mensal ou anual.

3.4 Transformando os dados

Concluído o processo de conhecimento dos dados, restaram 17 atributos, que novamente foram avaliados na questão dos valores ausentes conforme ilustração pela Figura 9.

De acordo com essa figura, alguns atributos possuem até 32,5 % de valores ausentes, sendo os demais com um valor muito inferior a 1 %, que não oferece o risco criar um conjunto de dados não representativo do problema real em uma dinâmica de substituição de valores ausentes.

Figura 9 – Valores ausentes por atributos



Fonte: Elaborada pelo autor.

Em alguns algoritmos de aprendizado de máquina, a presença de valores ausentes pode afetar o funcionamento de forma parcial ou total, reduzindo a acurácia preditiva do modelo induzido (CLAVEL; MERCERON; ESCARGUEL, 2014). Por isso, é necessário buscar formas de inserir um valor, que pode ser:

- um valor que indique que ele substitui um valor ausente;
- o valor médio, a mediana ou o valor de maior frequência;
- um valor estimado por alguma técnica de estimação de valores (ALLISON, 2002).

A solução adotada nos experimentos realizados neste trabalho para a inserção de dados nos atributos com valores ausentes foi a de calcular a média ou moda global de cada atributo e substituir pelo resultado.

Concluído o processo de substituição dos valores ausentes, o passo seguinte foi o de converter os atributos do tipo Qualitativo para Quantitativo, onde existem diversas abordagens para esse tipo de conversão (FACELI *et al.*, 2011), sendo adotada a substituição dos dados qualitativos de cada atributo por pseudo atributos inteiros, pois a maioria dos atributos podem ser representados por distâncias entre os mesmos justificando a codificação.

Os demais atributos foram mantidos como inteiros ou valor real, mas para os atributos que são do tipo Quantitativos discretos relacionados a datas foram extraídos apenas o dia.

Desse modo todos os atributos passaram a ser representados por números o que tornou possível calcular algumas propriedades de inferência estatística tais como a média, desvio padrão e quantil dos atributos conforme Tabela 4.

Tabela 4 – Análise estatística do conjunto de dados.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
mean	24.35	13.54	22.16	1.41e-05	1493.64	11.20	29.86	1.78	14.08	1.95e+05	2.50	4.39e-04	0.22	0.11	1.0	23.95	10.43
std	37.22	8.33	11.82	3.76e-03	682.38	6.33	7.65	0.55	7.54	6.20e+06	2.34	2.09e-02	0.42	0.65	0.0	10.03	17.71
min	0.00	0.00	0.00	0.00e+00	0.00	0.00	0.00	0.00	0.00	0.00e+00	0.00	0.00e+00	0.00	0.00	1.0	1.00	0.00
25%	0.00	8.00	12.00	0.00e+00	960.00	7.00	24.00	2.00	7.00	1.80e+04	1.00	0.00e+00	0.00	0.00	1.0	15.00	1.00
50%	0.00	12.00	19.00	0.00e+00	1739.00	11.00	35.00	2.00	19.00	7.83e+04	1.00	0.00e+00	0.00	0.00	1.0	23.00	3.00
75%	42.00	20.00	37.00	0.00e+00	2029.00	13.00	35.00	2.00	19.00	1.64e+05	3.00	0.00e+00	0.00	0.00	1.0	33.00	12.00
max	**127.00**	30.00	52.00	1.00e+00	2447.00	31.00	40.00	2.00	28.00	1.67e+09	9.00	1.00e+00	1.00	4.00	1.0	45.00	**127.00**

Fonte: Dados da pesquisa.

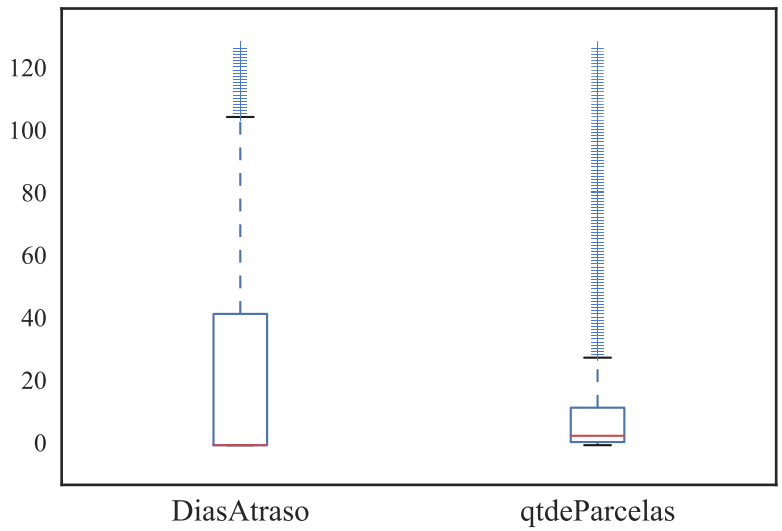
Outra validação muito importante associada a qualidade dos dados é a verificação da presença de *outliers*. Os *outliers* são atributos com valores máximos distantes da média ou do terceiro quantil, conforme Tabela 4, os atributos *qtdeParcelas* e *DiasAtraso* apresentaram valores máximos distantes, isso pode ser confirmado através do gráfico de boxplot ilustrado na Figura 10.

Os valores *outliers* podem influenciar de forma negativa alguns algoritmos de aprendizado de máquina na questão das fronteiras de decisão uma vez que podem se ajustar a esses dados como também podem acabar ignorando tais informações.

Algumas literaturas recomendam a substituição dos *outliers* por algum valor aproximado estimado pela técnica de EM ou até pelo valor médio, mas isso pode ocasionar uma perda de informação visto que alguns algoritmos são robustos suficientes para trabalhar com fronteiras como é caso do SVM.

Outra medida a ser tomada é utilizar uma técnica de normalização, com o objetivo de ajustar a escala dos valores de cada atributo de forma independente em um mesmo intervalo, onde podem ser utilizadas diversos tipos de técnicas de normalização. A mais comum é a normalização pelo desvio padrão que, embora não garanta a remoção de *outliers*, consegue deixar a distribuição

Figura 10 – Boxplot dos atributos que apresentaram outliers



Fonte: Elaborada pelo autor.

dos dados muito semelhante a uma distribuição gaussiana tendendo a média. Isso permite ou beneficia o funcionamento de vários algoritmos de aprendizado de máquina, como, por exemplo, as máquinas de vetores de suporte com kernel gaussiano e os algoritmos de regressão logística utilizando os regularizadores L1 e L2.

Esses dois casos assumem que todos os atributos estão centralizados e tendem a 0 e possuem variação na mesma ordem. Se um atributo apresentar uma variância com ordens de magnitude maiores que outras, ele pode influenciar a função objetiva e tornar o estimador incapaz de aprender e se ajustar aos demais atributos de forma correta conforme o esperado.

Existem, no entanto, técnicas de normalização que são robustas e conseguem tratar a questão dos *outliers* utilizando os valores de percentis ou definidos a priori em um intervalo de 0 a 1. A técnica utilizada para normalizar os dados do trabalho foi a de transformação por quantil desenvolvida na biblioteca Sklearn (PEDREGOSA *et al.*, 2011) que foi baseada no conceito de *inverse normal transformations* (INTs) (BEASLEY; ERICKSON; ALLISON, 2009), que consiste em calcular o posto de cada atributo e realizar uma interpolação para aplicar em uma função de densidade cumulativa para cada atributo.

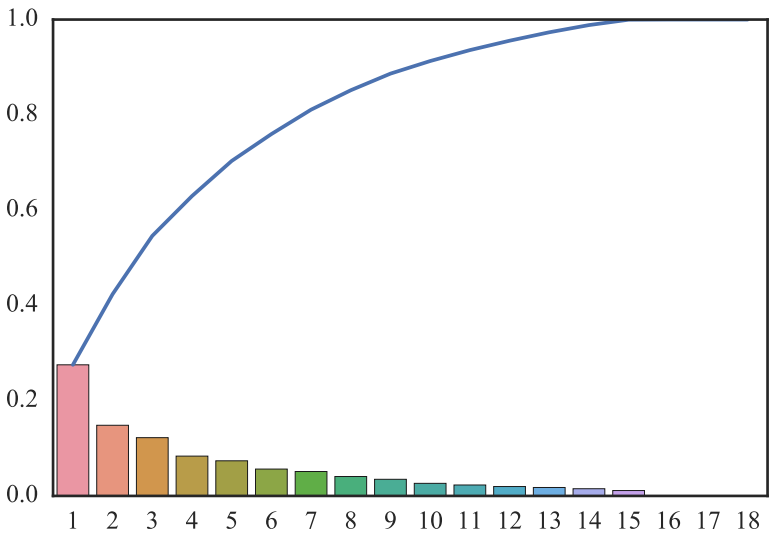
Essa função tem um papel de delimitar uma fronteira com escalas aceitáveis e os *outliers* presentes nos atributos serão mapeados para os limites da distribuição. Mas esta transformação não é linear e pode gerar em alguns casos distorções nas correlações lineares entre as variáveis medidas na mesma escala, mesmo assim a técnica torna as variáveis medidas em diferentes escalas mais diretamente comparáveis e de certo modo os dados normalizados ainda representam a generalização da amostra calculada.

Após calculada e aplicada a normalização dos dados, foi calculado o PCA com o

objetivo de visualizar inicialmente o comportamento dos dados por meio da projeção dos 3 primeiros componentes principais. De acordo com a técnica, os componentes representam a variância de apenas 52% inviabilizando uma redução de dimensionalidade que conforme a Figura 11 foi possível observar que somente no décimo quinto componente essa variância se torna representativa.

Desse modo foram mantidos os 17 atributos do conjunto de dados descartando a possibilidade de redução de dimensionalidade pela técnica de PCA evitando perda de informações.

Figura 11 – Variância dos componentes principais



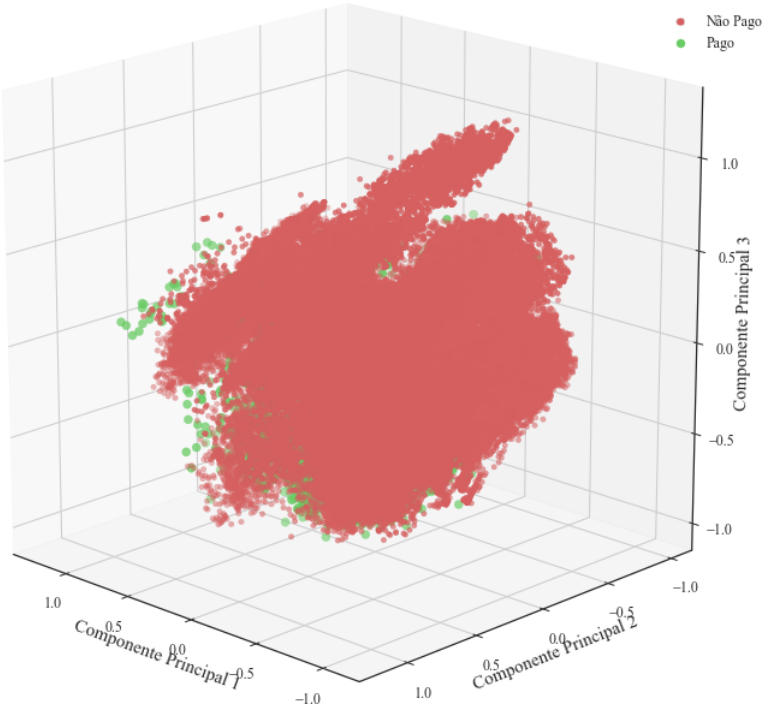
Fonte: Elaborada pelo autor.

A técnica de PCA pode ser utilizada não só como uma forma de redução de dimensionalidade como também uma forma de projetar os objetos em um espaço R2 ou até R3. Utilizando os componentes principais calculados de modo a visualizar as direções de variação dos objetos presentes no conjunto de dados é possível validar uma se existe uma linearidade.

Para visualizar o PCA, foram projetados os três primeiros componentes principais em um gráfico de *scatterplot* 3D para verificar se existe uma linearidade nos dados entre os componentes e o atributo resposta. Conforme Figura 12, foi possível verificar que dentre a representação de 52% de variância dos três primeiros componentes projetados, os objetos estão dispostos de forma não lineares e com grande sobreposição.

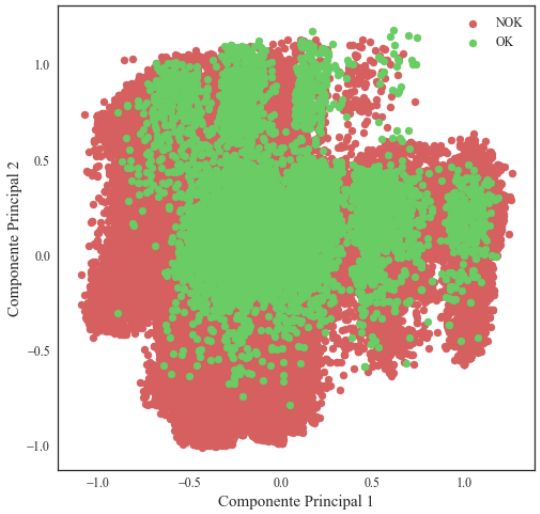
Na Figura13 a visualização dos objetos projetados nos dois primeiros componentes principais é um pouco mais clara, mas demonstra que a disposição do espaço de certo modo pode ter impacto nas técnicas de aprendizado de máquina que não estão preparadas para trabalhar com tal disposição por exemplo um rede neural simples ou Naive Bayes.

Figura 12 – Projeção dos três primeiros componentes principais dos dados já normalizados.



Fonte: Elaborada pelo autor.

Figura 13 – Projeção dos dois primeiros componentes principais dos dados já normalizados.



Fonte: Elaborada pelo autor.

3.5 Reamostragem dos dados

Conforme observado nos trabalhos pesquisados e descritos no Capítulo 2, alguns conjuntos de dados apresentam problemas de desbalanceamento nos atributos respostas. Em sua grande maioria as classes estão desproporcionais em até dez vezes, conforme trabalho (OLIVEIRA; SERPA, 2013), trata-se de um problema recorrente na grande maioria dos casos com tratamentos

de conjunto de dados bancários.

Em quase todos os processos do ciclo de crédito, foi possível observar a questão do desbalanceamento de classes, sendo que em alguns trabalhos como (XIE *et al.*, 2009) e (FARQUAD; RAVI; RAJU, 2014), foram adotadas técnicas para realizar uma reamostragem com a imputação de dados cópias ou sintéticos com o objetivo de igualar ou aproximar a quantidade da classe minoritária a majoritária, processo conhecido como *oversample*. O contrário, diminuindo a quantidade de dados da classe majoritária igualando a classe minoritária é conhecido como *undersample*.

Já no trabalho (MACHADO, 2016) essa questão foi ignorada, o que pode ter afetado o resultado da técnica de regressão logística, que conforme matriz de confusão presente na pesquisa, o modelo não conseguiu ajustar-se as classes minoritárias.

Geralmente os casos de recuperação de crédito e cobrança a presença de pagadores pode ser proporcionalmente menor, conforme observado no conjunto de dados do trabalho, a empresa de cobrança consegue recuperar em média 5% mensalmente, ou seja, um valor relativamente baixo para expressar significativamente o espaço de busca, afetando negativamente os algoritmos de aprendizado de máquina (BROWN; MUES, 2012).

Desse modo, conforme trabalhos pesquisados algumas técnicas de balanceamento de classes foram adotadas. A primeira encontrada foi a de *over-sample*, que consiste em copiar os objetos da classe minoritária em uma quantidade próxima a majoritária.

A outra técnica utilizada foi a de *under-sample* onde é aplicada uma remoção aleatória de objetos da classe majoritária até atingir uma quantidade próxima da minoritária.

Por último a destacar é a técnica de implantação de dados sintéticos chamada SMOTE (CHAWLA *et al.*, 2002). Esse tipo de abordagem foi amplamente utilizada para tratar dados financeiros desbalanceados conforme os trabalhos (Baesens, B. , Rösch, D. and Scheule, 2017) e (GARCÍA; MARQUÉS; SÁNCHEZ, 2012). Quando associada a outras técnicas apresentou também resultados satisfatórios em diversos casos de classes desbalanceadas conforme trabalho (BATISTA; PRATI; MONARD, 2004).

A técnica SMOTE (*Synthetic minority over-sampling technique*) é um método de *over-sampling* que difere dos métodos tradicionais onde somente replicam os dados de forma aleatória e consequentemente reproduzem um certo nível de redundância. Sua forma de replicação consiste em realizar uma interpolação entre dois exemplos próximos da classe minoritária, utilizando k-Vizinhos mais próximos. Em seguida alguns dos vizinhos são selecionados e um novo exemplo é gerado novamente através de interpolação. Desse modo, a técnica visa contribuir para que algumas técnicas de aprendizado de máquina "visualizem"as classes de forma semelhante reduzindo a taxa de falsos positivos.

Ao resolver a questão dos dados desbalanceados, o conjunto de dados está preparado para ser utilizado nos algoritmos de aprendizado de máquina propostos nesse trabalho. Sua

utilização consiste no processo de treino dessas técnicas utilizando as vezes técnicas de validação cruzada, *bootstraping* ou até mesmo separando uma amostra de 70% para treino e 30% para teste conforme observado nos trabalhos pesquisados. O uso dessas técnicas são de extrema importância para validar o desempenho dos classificadores em diferentes intervalos do conjunto de dados evitando problemas de *underfitting* ou *overfitting*.

CAPÍTULO

4

ALGORITMOS DE APRENDIZADO SUPERVISIONADO

Nesse capítulo, serão abordados os assuntos referentes a aos principais algoritmos de aprendizado de máquina utilizados nos trabalhos relacionados, suas principais aplicações em modelos de risco de crédito como também compara-las de modo a obter a melhor técnica para o desenvolvimento de um modelo de *Collection Score*.

Para realizar essas comparações, serão utilizados alguns cálculos e medidas de avaliação de erro durante o treinamento dos algoritmos de aprendizado de máquina supervisionados. Essas medidas são utilizadas para validar o desempenho dos modelos gerados por cada algoritmo e, posteriormente, para comparar o desempenho de diferentes classificadores.

Além das medidas de avaliação e comparação, outros temas como os tipos de treino, ajustes nos hiper-parâmetros dos algoritmos de aprendizado de máquina e motivo da escolha da abordagem de cada algoritmos em relação à área financeira serão abordados nesse capítulo.

4.1 Avaliação e modelagem do risco

Conforme trabalhos pesquisados no Capitulo 2, o modelo proposto por esse trabalho, se enquadra na área financeira chamada de Avaliação e Modelagem de Risco. Essa área em questão tem como seu principal objetivo quantificar o risco de forma especulativa, utilizando quando possível os dados históricos, sendo que esse risco está relacionado a rentabilidade, exposição (máximo de prejuízo suportado) e perda do portfólio a ser avaliado.

Para que esses modelos sejam desenvolvidos, é necessário cumprir um ciclo semelhante ao processo de mineração de dados, onde existem diversas etapas tais como a etapa de definição do problema, coleta dos dados, desenvolvimento do modelo, implementação, aplicação e acompanhamento.

As primeiras etapas do ciclo já foram descritas no Capítulo 3 descrevendo todo processo que envolve a definição e coleta de dados como também o preparo desses dados para a etapa de implementação, que será exposta nesse capítulo. Nessa etapa é desenvolvido o modelo que pode ser categorizado em três tipos: Modelos Financeiros (Estrutural, Fluxo de Caixa e Mercado Implicado), Modelos baseados em dados empíricos (Estatísticos e Aprendizado de Máquina) e Modelos Especialistas (Regras e Tributação) (GESTEL; BAESENS, 2009).

A categoria que será desenvolvida nesse trabalho é a de modelos baseados em dados empíricos. Esses modelos são fundamentados nas áreas de estatística e aprendizado de máquina e utilizam técnicas como a Regressão Logística, Redes Neurais, SVM, K-NN e Árvores de Decisão dentre outras técnicas.

Em relação os demais tipos de categorias de modelos, não serão descritas nesse trabalho e estão disponíveis no trabalho de (GESTEL; BAESENS, 2009).

4.1.1 Modelos baseados em estatística

Grande maioria dos trabalhos relacionados a modelagem para *scores* são desenvolvidos utilizando técnicas de regressão logística como é possível observar no Capitulo 2.

A regressão logística faz parte de um grupo de modelos chamado Modelo Linear Generalizado (MLG). Sua técnica basicamente consiste em examinar como um atributo alvo ou variável resposta é explicado pelos atributos de entrada, ou variáveis independentes, resultando em um conjunto de fatores chamados coeficientes de regressão ou conhecido por *logit*. Se o valor p de um coeficiente for menor que o nível de significância especificado, por exemplo 0,05 a relação entre o atributo e a resposta é estatisticamente significante. Desse modo, quanto mais atributos com valor de p próximos de 0, melhor será o grau de ajuste do modelo, o que faz da técnica uma excelente ferramenta para estudo e análise de atributos.

Conforme Tabela 5, quase todos os atributos apresentaram um valor p significante, todos destacados com ***, os valores foram calculados em um intervalo de confiança de 0,90, os demais atributos como *co_estado* e *co_bloqueia_notificacao* não apresentaram relevância. Para o modelo já coeficiente do atributo *cs_status* não foi calculado pelo modelo.

Mesmo com a possibilidade de saber qual atributo é relevante para utilização na construção de modelos, ainda assim não se pode avaliar a qualidade do modelo por meio dessa condição, pois é necessário também avaliar o desempenho de classificação através de medidas de avaliação.

Grande parte dos trabalhos pesquisados na área de análise e risco de crédito, utilizam da técnica de regressão logística, pois além da fácil implementação a técnica apresenta uma transparência para com os dados, ou seja, não se trata de uma caixa preta permitindo a realização de análises para cada atributo utilizado em um modelo e a extração de regras.

Tabela 5 – Coeficientes da Regressão Logística.

	Variável dependente:
	target
DiasAtraso	0.045*** (0.041, 0.048)
DtVencPar	0.047*** (0.043, 0.052)
cf_id	0.200*** (0.193, 0.206)
co_bloqueia_notificacao	0.057 (−0.325, 0.439)
co_cidade	0.021*** (0.017, 0.026)
co_data_entrada	−0.029*** (−0.033, −0.024)
co_estado	0.015*** (0.009, 0.021)
co_recomenda_debito	0.139*** (0.134, 0.144)
co_tipo_carteira	−0.041*** (−0.049, −0.034)
co_total	−0.092*** (−0.097, −0.087)
cr_id	−0.210*** (−0.216, −0.203)
cs_altera_status_devedor	−0.071** (−0.125, −0.016)
cs_bloqueia_lista	−0.060*** (−0.064, −0.057)
cs_excluido	0.021*** (0.014, 0.029)
cs_status	
en_id	−0.112*** (−0.117, −0.106)
qtdeParcelas	−0.013*** (−0.019, −0.007)
Constant	0.050*** (0.040, 0.060)
Observations	98,965
R^2	0.113
Adjusted R^2	0.113
Residual Std. Error	0.232 (df = 98948)
F Statistic	790.102*** (df = 16; 98948)
Nota:	*p<0.1; **p<0.05; ***p<0.01

Fonte: Dados da pesquisa.

4.1.2 Modelos baseados em procura

Conforme os trabalhos pesquisados e relacionados a classificação de problemas financeiros descritos no Capítulo 2, em sua grande maioria utilizaram árvores de decisão ou técnicas similares de divisão e conquista (*greedy algorithm*). Os trabalhos que obtiveram resultados satisfatórios utilizaram CART e Random Forest, ambos são baseados em árvores de decisão, porém com diferentes tipos de abordagens, sendo CART uma técnica de árvore de decisão e RandomForest utiliza-se de árvores porém com uma abordagem com o conceito de *Ensemble learning* combinando outras técnicas.

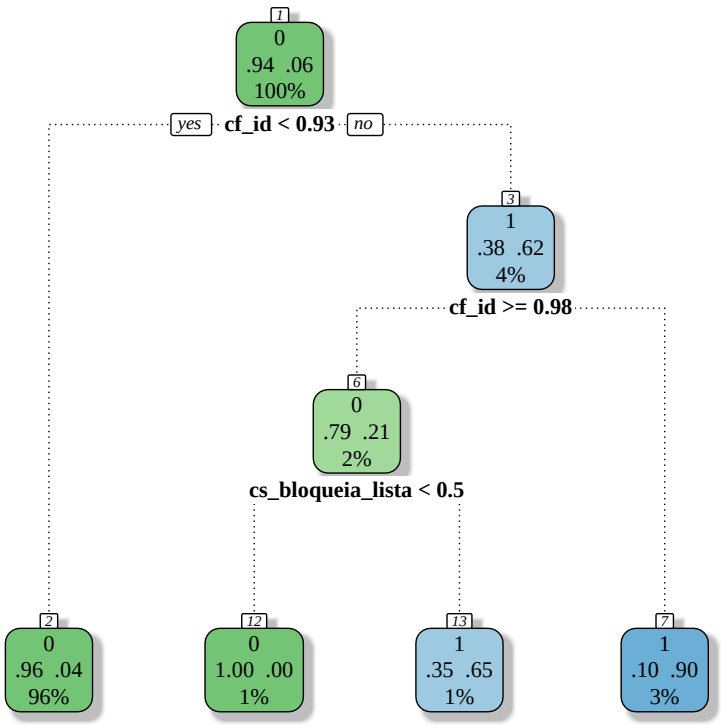
Ensemble learning constitui-se como uma vertente do Aprendizado de Máquina que trabalha com a combinação de modelos. Por intermédio da utilização de *ensembles* é possível transformar um esquema de aprendizado relativamente fraco em um mais robusto (WITTEN; FRANK; HALL, 2011). Esse resultado só é possível quando existe uma combinação desses classificadores fracos visando obter um classificador forte, desde que haja diversidade entre os classificadores (Ponti Jr, 2011).

O método *Random Forest* é constituído por uma combinação de árvores de decisão, sendo que sua perspectiva de amostragem aleatória e estratégias *ensemble* utilizadas no RF permitem

que este consiga realizar predições precisas, bem como melhores generalizações (QI, 2012). O resultado do modelo é dado pela soma dos votos de cada classificador encontrado, sendo que para um dado objeto a classe atribuída é a que obtiver a maior quantidade de votos. Desta forma, de acordo com o princípio de *ensemble*, o *Random Forest* constitui o "*strong learner*"gerado a partir da combinação das árvores de decisão intermediárias que constituem os "*weak learners*".

Para comparar o desempenho do classificador anterior foi utilizado também a técnica de árvores de decisão - CART (MITCHELL, 1997) diferentemente do método *essemble* do qual é construído todo conceito do RF, as árvores de decisão contam apenas com um classificador e uma medida *gini* ou entre outras, onde uma vez definida, inicia-se o processo que calcula um valor índice de cada atributo do conjunto de dados, com o objetivo de avaliar qual é o melhor atributo para iniciar a ramificação da árvore, que é construída por um algoritmo guloso.

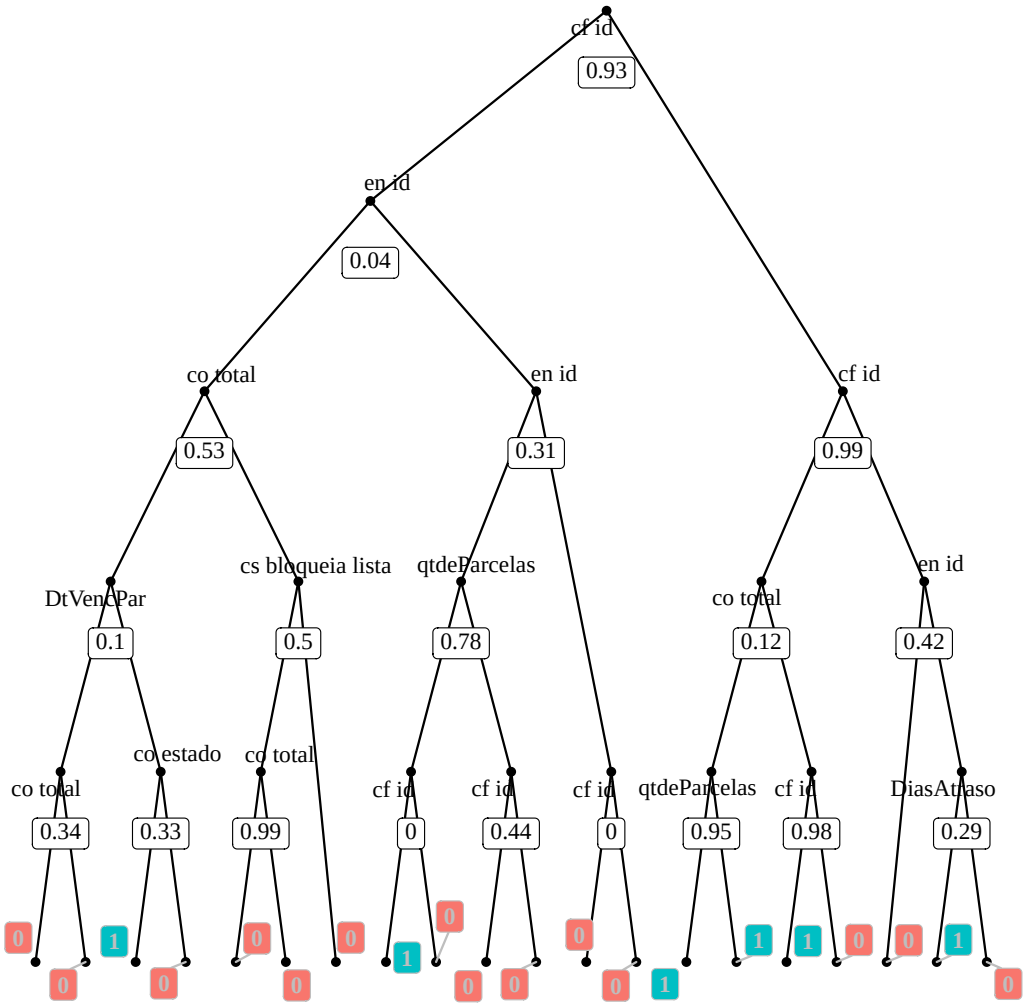
Figura 14 – Árvore - CART



Fonte: Dados da pesquisa.

Conforme Figura 14, é possível visualizar a estrutura da árvore de decisão, que de certo modo revelaria como funciona a regra de negócio da empresa de cobrança, ou seja, quais possíveis atributos são importantes para encontrar bons e maus pagadores. Mas a técnica não conseguiu explorar os demais atributos presentes, realizando as divisões das classes com apenas 3 nós sendo eles dois atributos, o *cf_id* que está relacionado a fase da cobrança e depois o atributo *cs_bloqueia_lista* que está relacionado a opção de bloqueio informado pelo banco e as respostas 1 para Pago e 0 para Não Pago.

Figura 15 – Maior Árvore - RandomForest



Fonte: Dados da pesquisa.

A técnica de RandomForest também possui uma representação gráfica, porém a técnica gera mais de uma árvore, sendo necessário ilustrar todas, ou utilizar alguma técnica para analisar de forma sintética os atributos.

O modelo gerou 20 árvores, para ilustrar a técnica foi utilizada a árvore maior, conforme a Figura 15. Desse modo foi possível verificar que a técnica conseguiu trabalhar com os demais atributos um pouco melhor com relação à técnica CART, pois além dos já encontrados por ela, novos foram utilizados como o *en_id* que é o banco, *co_total* é o valor da dívida, *co_estado* a unidade federativa, *qtde_parcelas* é o número de parcelas e *DtVencPar* em qual dia do mês que elas vencem e as respostas 1 para Pago e 0 para Não Pago.

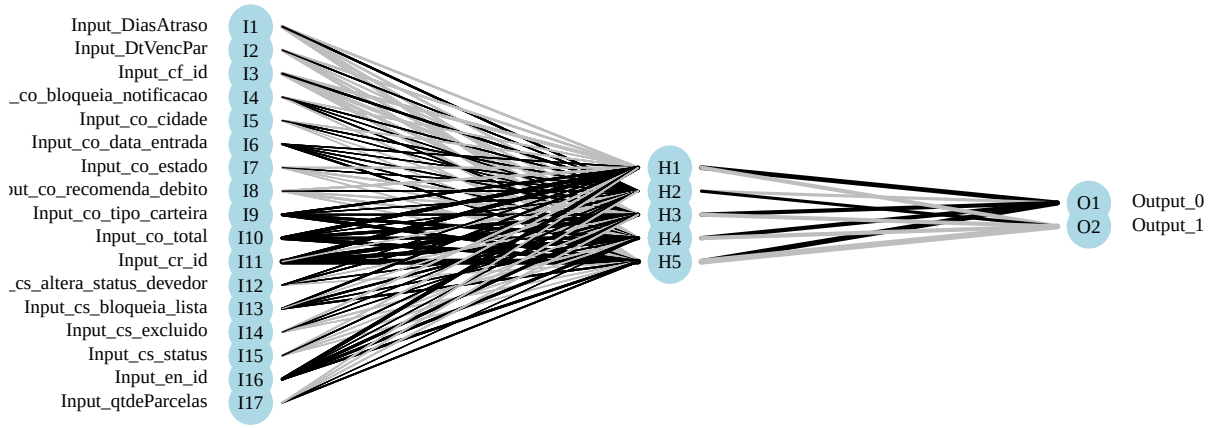
Lembrando que para melhor análise é necessário representar as demais árvores, sendo assim no próximo capítulo será descrita outra abordagem para avaliar esse modelo. É importante ressaltar também que as estruturas construídas representam o cenário de cobrança da empresa

estudada e não está relacionada ao contexto geral de todas as cobranças do país.

4.1.3 Modelos baseados em otimização

Outra abordagem muito utilizada para soluções de problemas de classificação é o uso de redes neurais artificias ou RNA. Diferentemente dos algoritmos de procura como RandomForest ou CART que iniciam sempre com uma entrada, ou nó e são construídos conforme iterações, a estrutura da rede neural precisa ser definida antes da execução dos cálculos. Os nós que estão presentes nas técnicas de busca, no conceito de redes neurais são representados por neurônios e cada ligação ou sinapse possui um valor chamado peso. Essa estrutura é constituída por três camadas, a primeira chamada de entrada, possui um neurônio para cada atributo presente no conjunto de dados. A seguinte camada chamada de intermediária é responsável pela construção da rede neural e faz ligações com outras camadas intermediárias ou com a última camada chamada de saída, onde cada possível resposta é representada por um (regressão) ou mais neurônios (classificação).

Figura 16 – Representação Rede Neural MLP



Fonte: Dados da pesquisa.

O treino de uma RNA é como a execução de algoritmos de otimização, pois se baseia em funções para minimizar o erro entre as respostas da rede e os rótulos esperados, uma vez inserido a informação por meio do neurônios de entrada, os demais seguintes são ativados por outro tipo de função chamada de ativação. Semelhante ao processo biológico de ativações de sinapses, o processo da rede é executado até chegar a camada de saída. Caso o valor deduzido pela rede esteja errado, é executada uma função para corrigir os pesos e o processo reinicia até que a resposta fique correta. Esses processos são conhecidos como *"forward"* e o de correção *"backward"*, formando o processo de treino conhecido como *backpropagation*.

Existem basicamente dois tipos de RNA, a Rede Neural ou Perceptron e a Rede Neural Multi-Camada, conhecida como MLP. A rede com perceptron é utilizada para solução de problemas lineares, e a MLP para problemas não lineares, sendo necessário definir o tamanho

e a quantidade de camadas ocultas antes do treino da rede MLP, de acordo com (MASTERS, 1994), o correto é iniciar sempre o treino com um número menor de neurônios e camadas, e sucessivamente aumentar de acordo com o nível de acerto, até o ponto em que a taxa de acerto torna-se invariante, evitando assim problemas como *overfiting* ou ótimos locais. Nesse trabalho, foi definida uma rede MLP com uma camada intermediária e definido 20 neurônios, porém a ferramenta de otimização encontrou um melhor resultado com apenas 5 neurônios, comprovando conforme a literatura, não são necessários muitos neurônios para a solução de um problema complexo (MELLO; PONTI, 2018).

4.2 Medidas de avaliação

Os classificadores foram avaliados através das medidas clássicas de desempenho, são elas: Acurácia (ACC), Especificidade (ESP), Sensibilidade (SEN) e Kappa. Todas estas medidas são calculadas a partir da matriz de confusão resultante de cada classificador. Esta matriz tem geralmente duas linhas e duas colunas que informam o número de falsos positivos (FP), falsos negativos (FN), verdadeiros positivos (VP), e verdadeiros negativos (VN) e também será exposta nesse trabalho para melhor expressão dos resultados e posteriormente servir de base para novos trabalhos utilizando outras medidas de avaliação.

Embora esse trabalho forneça os resultados para estas medidas, para eleger o melhor classificador será considerada apenas uma medida, a medida Kappa (COHEN, 1960), que é a medida mais apropriada para comparar classificadores distintos. Kappa mede o quão perto os casos rotulados pelos classificadores estão dos valores reais. Assim, o índice Kappa de um classificador é propriamente comparável a outros kappas de classificadores para a mesma tarefa de classificação (PRATI; BATISTA; SILVA, 2015).

CAPÍTULO

5

EXPERIMENTOS E RESULTADOS

Todas as técnicas foram treinadas com os dados do segundo semestre de 2016 e primeiro semestre de 2017, selecionando uma amostra de forma aleatória de 70% para treino e o restante para teste. Além da amostragem foi utilizada uma técnica de validação cruzada K-Fold, com 10 folds com o objetivo de validar possíveis ruídos ou algum tipo de viés gerado por algum evento econômico.

Para executar as técnicas de aprendizado de máquina, foi utilizada a linguagem R com as bibliotecas *caret* (KUHN; KUHN, 2009) (Para treinos e validações), *GLM* (CHAMBERS; HASTIE, 1991) para regressão logística, *RPart* (THERNEAU; ATKINSON; RIPLEY, 2011) para árvores de decisão CART, *RandomForest* (LIAW; WIENER, 2002), para a técnica de RandomForest e a *RSNSS* (BERGMEIR; BENÍTEZ, 2012) para redes neurais MLP.

5.1 Ajuste de hiperparâmetros

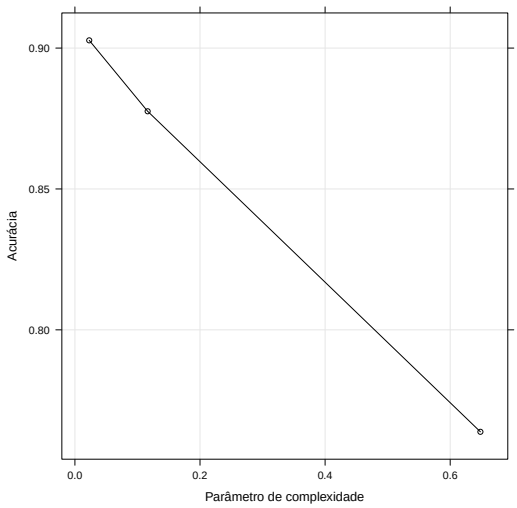
Em todas as técnicas conforme observado no Capitulo 4, foram realizadas otimizações nos valores dos hiperparâmetros pelos algoritmos automaticamente pela biblioteca Caret, utilizando a técnica de *Grid Search*.

Iniciando pela técnica de árvore de decisão, onde o melhor valor de complexidade ou CP foi de 0,02 apresentando um valor Kappa de 0,79 e acurácia de 0,90 conforme a Figura 17.

Apesar do resultado satisfatório, a estrutura da árvore de regressão CART não foi muito representativa para o ponto de vista da empresa de cobrança, pois conforme a Figura 14, apenas dois atributos foram selecionados, muito pouco para permitir a extração de regras do modelo.

Outra técnica que contou com o recurso de otimização foi a de RandomForest, que apresentou um resultado melhor que a técnica anterior, tanto na questão do acerto como também na indução da árvore de decisão. O parâmetro otimizado foi o mty (Número de atributos aleatoriamente reamostrados como candidatos para cada divisão) sendo 9 o melhor valor representando

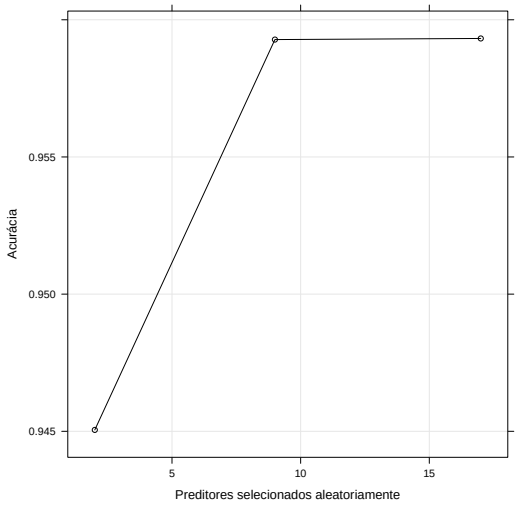
Figura 17 – Otimização CART



Fonte: Dados da pesquisa.

um valor Kappa de 0,92 e uma acurácia de 0,96 conforme Figura 18, foi de fato a técnica que apresentou o melhor resultado.

Figura 18 – Otimização - Random Forest



Fonte: Dados da pesquisa.

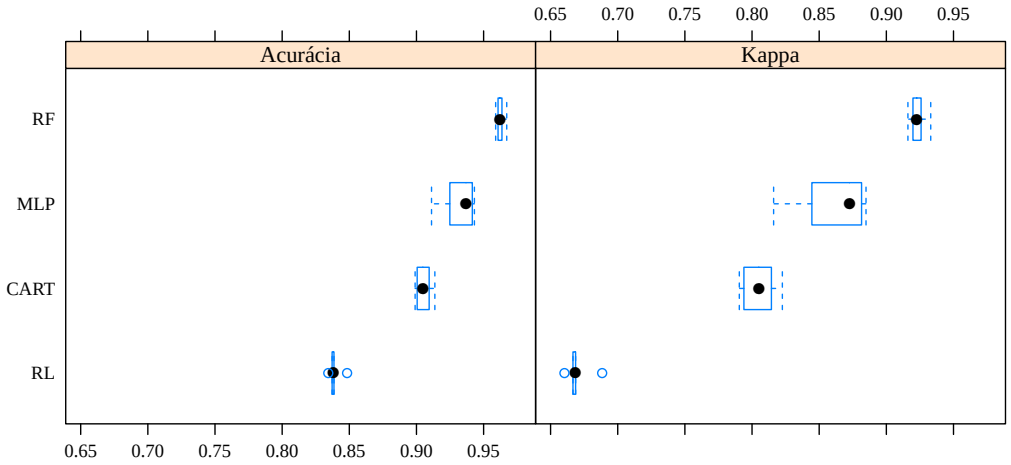
Por último, as redes neurais multicamadas, redes MLP, também tiveram seus hiperparâmetros otimizados, apesar da necessidade de definir um número mínimo de neurônios e camadas ocultas no caso 20, o melhor resultado apresentado foi o de 5 neurônios com um valor Kappa de 0,85 e acurácia de 0,93.

A técnica de regressão logística utilizada em grande escala por trabalhos na área de risco de crédito conforme pesquisado e descrito no Capítulo 2, apresentou um resultado muito inferior

às técnicas anteriores, não sofreu nenhum tipo de otimização e o valor de Kappa foi 0,67 a acurácia de 0,83.

Todos os modelos foram submetidos a técnica de validação cruzada conforme citado anteriormente, a Figura 19 ilustra todos os 10 resultado através de bloxplots.

Figura 19 – Boxplot das validações cruzadas dos classificadores



Fonte: Dados da pesquisa.

Tabela 6 – Resultado dos classificadores

Medidas	RF	B-RF	CART	B-CART	RL	B-RL	MLP	B-MLP
Acurácia	0,96	0,96	0,95	0,92	0,93	0,84	0,93	0,94
Sen	0,98	0,96	0,99	0,95	0,95	0,87	0,93	0,99
Espec	0,71	0,90	0,42	0,46	0,06	0,53	0,86	0,50
Kappa	0,72	0,72	0,54	0,40	0,11	0,24	0,59	0,61

Fonte: Dados da pesquisa.

Nota – Os classificadores foram induzidos pelos algoritmos RF (do inglês Random Forest), CART (do inglês Classification And Regression Tree), RL (Regressão Logística) e MLP (do Inglês Multi Layer Perceptron), todos aplicados nos dados de teste. A letra B antes das siglas significa que os dados foram balanceados pela técnica SMOTE

5.2 Validação dos modelos

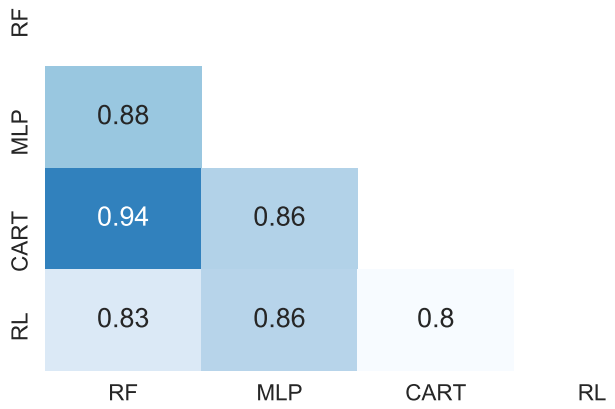
Após concluído o treino de todas as técnicas propostas, utilizou-se os 30% de amostra para validação dos modelos gerados, os resultados estão descritos na Tabela 6, onde ao analisá-la,

foi possível concluir que o classificador de melhor desempenho foi o Random Forest e em segundo lugar MLP ambos com as classes balanceadas pela técnica de SMOTE.

A o modelo MLP com os dados balanceados apresentou uma especificidade um pouco baixa com relação a técnica sem balanceamento, cerca de 0,50 indicando um percentual de erro maior em uma das classes. Esse percentual maior, se da pelo fato das classes estarem originalmente desbalanceadas e posteriormente foram balanceadas de forma sintética através da técnica SMOTE. Mesmo com as classes balanceadas sinteticamente, ainda existe uma dificuldade para o aprendizado de qualquer algoritmo supervisionado (HE; MA, 2013), onde nesses casos, os dados sintéticos nem sempre conseguem representar satisfatoriamente um espaço de busca favorável para os algoritmos, resultando em soluções baseadas em ótimos locais. O oposto aconteceu com a técnica de regressão logística, que apesar do índice Kappa estar baixo, conseguiu ter um desempenho melhor quando os dados estão balanceados.

As redes neurais MLP apesar de muitos considerarem uma "caixa-preta", existem trabalhos como (LU; SETIONO; LIU, 2017), (DUNHAM, 2003) e (DUCH; Adamczak Rafałand Grabczewski, 1996) que desenvolveram algoritmos para agrupar os neurônios que representam importância na resposta de determinados valores e extrair regras.

Figura 20 – Similaridade entre os resultado previstos para cada modelo



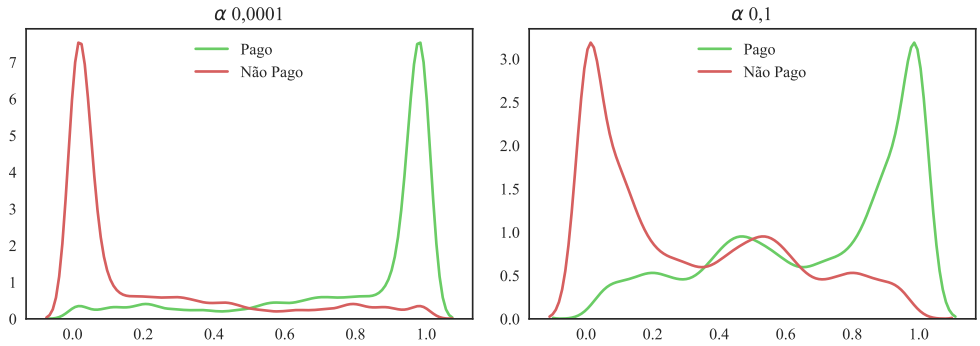
Fonte: Dados da pesquisa.

Tanto os modelos construídos pelas técnicas RandomForest e MLP apresentaram resultados muito próximos, foi realizada uma comparação entre os resultados previstos de cada modelos com o objetivo de validar a similaridade de cada resposta utilizando o cálculo do score de Similaridade de Jaccard (LABATUT; CHERIFI, 2011) e conforme Figura 20, ambos os modelos apresentaram valores de resposta próximos, que reforça o uso da reposta de um deles por exemplo o MLP, porém a extração das regras pode ser realizada ainda pelo RandomForest.

Apesar do modelo de redes neurais MLP anteriormente ter um papel de Classificação pois suas repostas são binárias ou categóricas, foi possível extrair também a probabilidade para

cada classe, com o principal objetivo de calcular o score de 0 a 5, conforme demonstrado na Figura 21, mas para que a técnica MLP ajustasse essa curva mais ampla conforme ilustrada, foi necessário utilizar o algorítimo de MLP da biblioteca Sklearn (GARRETA; MONCECCHI, 2013), que possui os regularizadores implementados. Ao aumentar o valor de α ou Regularização L2 para de 0,0001 para 0,1 foi possível diminuir a sensibilidade a alta variância, esse parâmetro também é utilizado para estabilizar os pesos quando existe uma alta correlação entre os atributos de entrada, ajudando a evitar superajustes (PERRIER, 2017).

Figura 21 – Gráfico de densidade dos resultados preditos pela técnica MLP



Fonte: Dados da pesquisa.

Nota – Ao alterar o valor de α para de 0,0001 para 0,1 foi possível obter uma curva mais ampla e consequentemente criar um limiar para cobrar contratos próximos dos recuperáveis

5.3 Interpretabilidade dos modelos

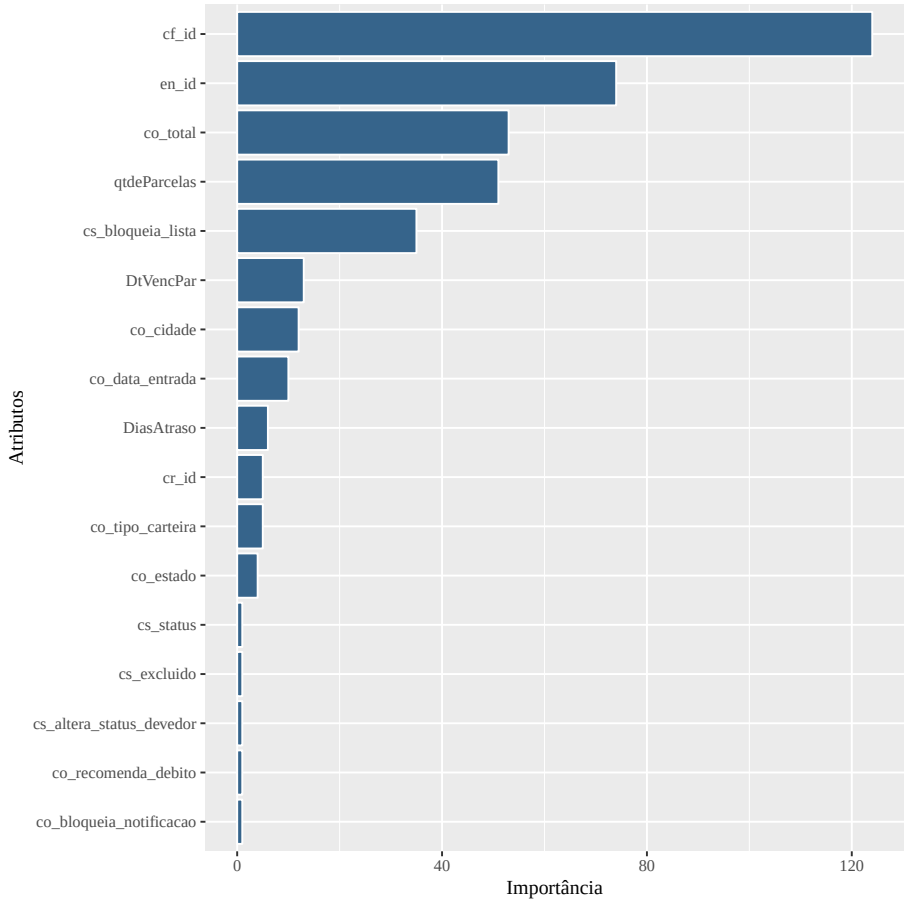
Uma das dificuldades de aceitação do uso de modelos gerados por técnicas de Aprendizado de Máquina é a interpretabilidade desses modelos, que permita entender de forma clara como eles geram cada uma de suas saídas. A interpretabilidade de modelos é cada vez mais exigida, principalmente em tarefas que usam dados pessoais.

Assim, outro ponto importante a se considerar além do melhor resultado é a capacidade de extrair regras dos modelos treinados pelas técnicas descritas anteriormente. O mais comum utilizado no mercado é validar a importância dos atributos através do valor p calculado pela regressão logística como também o *Odds Ratio*, ambos podem descriminar quais atributos são relevantes para o problema em questão.

Mas as técnicas de busca podem representar melhor o comportamento dos atributos para a classificação dos resultados, a técnica de RandomForest foi a que apresentou um melhor desempenho na construção das árvores, mas para uma possível extração de regras de negócio, pode-se recorrer a diversas abordagens, a mais comum pode ser a obervação das árvores geradas pela técnica, no caso desse trabalho, foram construídas 20 árvores, sendo que a maior árvore gerada está ilustrada na Figura 15.

Para realizar essa tarefa, seria necessário ilustrar as demais 19 árvores para tentar compreender o comportamento dos clientes e mesmo assim pode parecer algo extremamente complexo, ainda mais se a técnica gerar um número maior de árvores, esse tipo de abordagem acaba sendo inviável. Nesses casos, o que pode ser feito é contar quantas vezes os atributos apareceram nos nós das árvores, conforme a Figura 22, foi possível encontrar quais são os atributos mais importantes para o modelo descrever o comportamento das partições da classe negativa de não pagadores, representado pelo valor 0 e as positivas de pagadores, representadas pelo valor 1.

Figura 22 – Gráfico da importância dos atributos



Fonte: Dados da pesquisa.

Desse modo, foi possível obter o *score* para cada contrato a ser cobrado e possíveis novos contratos que estão com escores próximos dos recuperáveis, permitindo uma tomada de decisão rápida e ao mesmo tempo projetar através das ferramentas de BI cenários financeiros de cada carteira, como também para cada fase da cobrança e traçar ações melhores de recuperações para os casos isolados nas árvores de decisões construídas pela técnica de RandomForest e a capacidade de expressar a importância dos atributos.

CAPÍTULO

6

CONCLUSÃO

O principal desafio desse trabalho foi o de aplicar as técnicas de mineração de dados nas informações geradas pelo CRM de cobrança, com o objetivo de encontrar informações relevantes e que poderiam servir de apoio para tomadas de decisões e principalmente calcular um *score* para mapear bons pagadores.

Os dados coletados diretamente do CRM de cobrança da empresa, possuía quase todos os tipos de problemas possíveis elencados pela técnica de pré-processamento de dados, demandando um esforço maior no preparo dos dados para o treino posteriormente nas técnicas de aprendizado de máquina.

Ao realizar as etapas iniciais de mineração de dados, foi possível concluir o quão importante é o domínio da área de ciência de dados, pois se os dados não forem preparados corretamente antes de aplicá-los a qualquer tipo de algoritmo, as técnicas de aprendizado de máquinas podem sofrer de um efeito chamado de *Garbage-IN Garbage-OUT* GIGO (CABENA *et al.*, 1998). Se atributos considerados desconexos com o problema em questão, por um eventual descuido forem utilizados para o treino, consequentemente eles podem ser considerados importantes para a previsão do resultado gerando resultados indesejados ou que não representam uma solução para o problema em questão. Além da questão de seleção de atributos lixo, ainda existe um problema chamado escorregamento. Ao selecionar um atributo que dá informações sobre atributo alvo e não é previamente conhecido, onde passou a existir posteriormente (PROVOST; FAWCETT, 2013), o mesmo pode viciar o modelo e induzir a resposta.

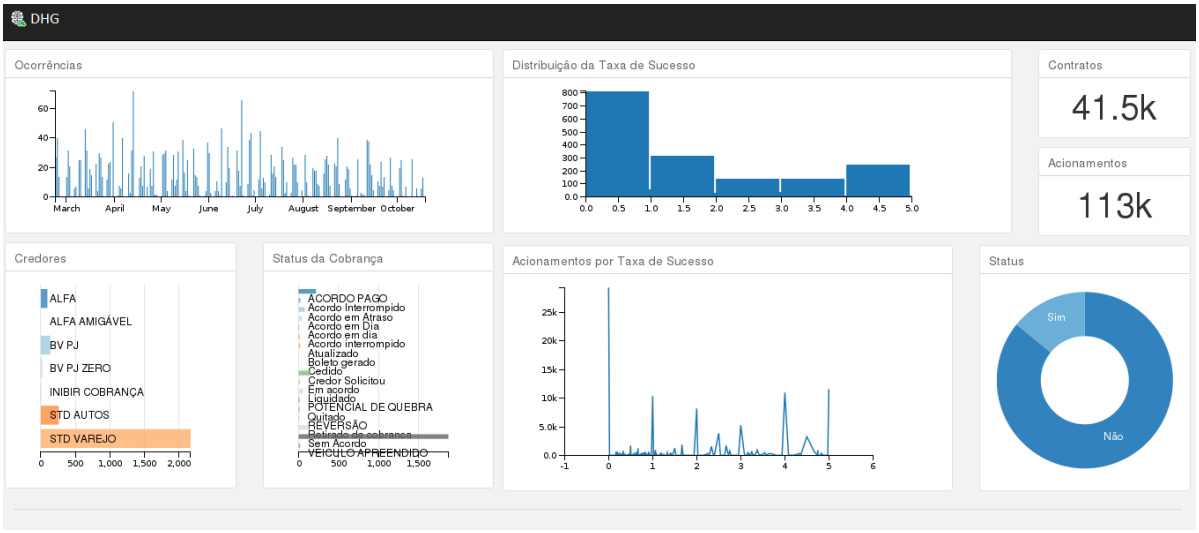
Outra questão importante abordada pelo trabalho foi sobre o problema de dados com classes desbalanceadas. De acordo com a pesquisa realizada, a grande maioria de modelos financeiros precisam lidar com classes desbalanceadas e em alguns casos, a performance é prejudicada resultando na não discriminação da classe minoritária, sendo necessário utilizar técnicas de reamostragem para balancear as classes.

Ao realizar a revisão bibliográfica foi possível concluir que grande parte dos modelos

de risco de crédito utilizam somente a técnica de regressão logística para classificação, nesse trabalho, além da regressão logística, foram empregadas outras técnicas de aprendizado de máquina, como árvore de decisão CART, RandomForest e Multilayer Perceptron. As novas técnicas implementadas também conseguiram resolver o problema relacionado ao *score* de cobrança que faz parte da área de risco de crédito. Elas não só demonstraram um melhor resultado como também possibilitaram a extração de regras de negócios representadas pelas árvores geradas pela técnica RandomForest. Com a técnica de Multilayer Perceptron, foi possível ajustar uma curva de classificação mais ampla permitindo a empresa de cobrança trabalhar com os casos próximos dos bons pagadores, resultando em um aumento gradativo da recuperação de cobrança.

As bibliotecas e linguagens open-source, permitiram a flexibilidade de desenvolver um módulo que foi integrado no CRM de cobrança da empresa e desenvolvido também um painel para visualização de Inteligência de Negócios, conforme Figura 23.

Figura 23 – Painel de Inteligência de Negócios



Fonte: Dados da pesquisa.

De acordo com o mercado financeiro de cobrança, a grande maioria das informações vitais sobre o processo de cobrança e negociação estão concentradas nos dados não-estruturados. Desde gravações telefônicas das negociações, registros de chat até canais em redes sociais, todos contribuem para uma grande massa de dados que podem explicar características do clientes devedores. Sendo assim, as futuras pesquisas e trabalhos estarão dentro do campo de aprendizado de máquina avançado e computação cognitiva para preparar e extrair informações.

O uso do aprendizado de máquina avançado para o treino de redes neurais avançadas, capazes de interpretar as gravações telefônicas será o principal objetivo de pesquisa dos próximos trabalhos. Para isso será necessário relacionar os conceitos de mineração de dados ao processo de big data para realizar o treino desses algoritmos, utilizando uma grande massa de dados,

será possível obter um sistema de reconhecimento de voz, conhecido como Automatic speech recognition (ASR).

Posteriormente à extração dos dados, será desenvolvida uma nova plataforma para o auxílio a tomada de decisão no momento da negociação da dívida com o cliente.

O curso da Universidade de São Paulo chamado Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria (MECAI), foi muito importante para concretização deste trabalho. A contribuição, ensino e orientação dos professores durante e depois das disciplinas, serviram de base sólida para a compreensão das técnicas empregadas. O principal foco do curso de mestrado profissional foi aplicar todo conhecimento adquirido na universidade e transferi-lo como recurso tecnológico ao mercado, que no caso desse trabalho foi desenvolvido um modelo matemático para o mercado financeiro, mais especificamente no segmento de recuperação de crédito.

REFERÊNCIAS

ALLISON, P. D. **Missing data**. [S.l.]: SAGE, 2002. 91 p. ISBN 9781452207902. Citado na página 43.

ANDERSON, R. **The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation**. Oxford University Press, 2007. Disponível em: <https://econpapers.repec.org/RePEc:oxp:obooks:9780199226405>. Citado na página 29.

Baesens, B. , Rösch, D. and Scheule, H. Low Default Portfolios. In: **Credit Risk Analytics**. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2017. p. 213–236. Disponível em: <http://doi.wiley.com/10.1002/9781119449560.ch8>. Citado na página 48.

BATISTA, G. E. A. P. A.; PRATI, R. C.; MONARD, M. C. A study of the behavior of several methods for balancing machine learning training data. **ACM SIGKDD Explorations Newsletter**, ACM, v. 6, n. 1, p. 20, jun 2004. ISSN 19310145. Disponível em: <http://portal.acm.org/citation.cfm?doid=1007730.1007735>. Citado na página 48.

BATTITI, R. Using mutual information for selecting features in supervised neural net learning. **IEEE Transactions on neural networks**, IEEE, v. 5, n. 4, p. 537–550, 1994. Citado na página 38.

BEASLEY, T. M.; ERICKSON, S.; ALLISON, D. B. Rank-based inverse normal transformations are increasingly used, but are they merited? **Behavior Genetics**, NIH Public Access, v. 39, n. 5, p. 580–595, sep 2009. ISSN 00018244. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/19526352http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2921808>. Citado na página 45.

BERGMEIR, C.; BENÍTEZ, J. M. Neural Networks in {R} Using the Stuttgart Neural Network Simulator: {RSNNS}. **Journal of Statistical Software**, v. 46, n. 7, p. 1–26, 2012. Disponível em: <http://www.jstatsoft.org/v46/i07/>. Citado na página 59.

BHAMBRI, V. Application of Data Mining in Banking Sector. **International Journal of Computer Science and Technology**, v. 2, n. 2, p. 4, 2011. Disponível em: <www.ijcst.com>. Citado na página 23.

BOTTOU, L. Online learning and stochastic approximations. **On-line learning in neural networks**, Cambridge Univ. Press, v. 17, n. 9, p. 142, 1998. Citado na página 37.

BROWN, I.; MUES, C. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. **Expert Systems with Applications**, Pergamon, v. 39, n. 3, p. 3446–3453, feb 2012. ISSN 09574174. Disponível em: <https://www.sciencedirect.com/science/article/pii/S095741741101342X>. Citado na página 48.

CABENA, P.; HADJINIAN, P.; STADLER, R.; VERHEES, J.; ZANASI, A. **Discovering data mining : from concept to implementation**. Prentice Hall, 1998. 195 p. ISBN 0137439806. Disponível em: <https://dl.acm.org/citation.cfm?id=270298>. Citado nas páginas 23, 26 e 65.

Castelar Pinheiro, A.; MOURA, A. Segmentation and the Use of Information in Brazilian Credit Markets. In: **Credit Reporting Systems and the International Economy**. [S.l.: s.n.], 2003. Citado na página 29.

CHAMBERS, J. M.; HASTIE, T. J. **Statistical Models in S**. [S.l.]: Chapman & Hall/CRC, 1991. 608 p. ISBN 0412053012. Citado na página 59.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. SMOTE: Synthetic minority over-sampling technique. **Journal of Artificial Intelligence Research**, v. 16, p. 321–357, 2002. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-0346586663{&}partnerID=40{&}md5=dfb419b8460388447758f9c7f8>. Citado nas páginas 31 e 48.

CLAVEL, J.; MERCERON, G.; ESCARGUEL, G. Missing Data Estimation in Morphometrics: How Much is Too Much? **Systematic Biology**, v. 63, n. 2, p. 203–218, mar 2014. ISSN 1076-836X. Disponível em: <http://dx.doi.org/10.1093/sysbio/syt100https://academic.oup.com/sysbio/article/63/2/203/1644797>. Citado nas páginas 38 e 43.

COHEN, J. A Coefficient of Agreement for Nominal Scales. **Educational and Psychological Measurement**, v. 20, n. 1, p. 37–46, apr 1960. ISSN 0013-1644. Disponível em: <http://journals.sagepub.com/doi/10.1177/001316446002000104>. Citado na página 57.

DASS, R. Data Mining in Banking and Finance: a Note for Bankers. **Indian Institute of Management Ahmedabad**, p. 1–15, 2006. Citado na página 23.

DERELIOUGLU, G.; GÜRGEN, F.; OKAY, N. A Neural Approach for SME's Credit Risk Analysis in Turkey. In: PERNER, P. (Ed.). **Machine Learning and Data Mining in Pattern Recognition**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. p. 749–759. ISBN 978-3-642-03070-3. Citado nas páginas 29 e 33.

DUCH, W.; Adamczak Rafałand Grabczewski, K. Extraction of logical rules from training data using backpropagation networks. In: . [S.l.: s.n.], 1996. Citado na página 62.

DUNHAM, M. H. **Data mining introductory and advanced topics**. [S.l.]: Prentice Hall/Pearson Education, 2003. 315 p. ISBN 0130888923. Citado na página 62.

FACELI, K.; LORENA, A. C.; GAMA, J.; CARVALHO, A. **Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina**. [S.l.: s.n.], 2011. ISBN 9788521618805. Citado nas páginas 25, 35, 36 e 44.

FARQUAD, M. A. H.; RAVI, V.; RAJU, S. B. Churn prediction using comprehensible support vector machine: An analytical CRM application. **Applied Soft Computing**, v. 19, p. 31–40, 2014. ISSN 1568-4946. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1568494614000507>. Citado nas páginas 31, 33, 34 e 48.

FERREIRA, R.; NETO, R.; MAURÍCIO, R.; SOBRINHO, B.; Marques Cavalcanti, A. **Estudo comparativo entre modelos de classificação para Behavior Scoring em procedimentos de análise de risco de crédito**. 2013. Citado nas páginas 30, 33 e 34.

GAMA, J.; CARVALHO, A. C. P. d. L.; FACELI, K.; LORENA, A. C.; OLIVEIRA, M.; OTHERS. **Extração de conhecimento de dados: data mining**. [S.l.]: Edições Silabo, 2015. Citado na página 39.

GARCÍA, V.; MARQUÉS, A. I.; SÁNCHEZ, J. S. Improving Risk Predictions by Preprocessing Imbalanced Credit Data. 2012. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1008.4071>. Citado na página 48.

GARRETA, R.; MONCECCHI, G. **Learning scikit-learn: Machine Learning in Python**. [s.n.], 2013. v. 12. 461–468 p. ISSN 0196-6553. ISBN 9781107671812. Disponível em: <http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html>. Citado na página 63.

GEMAN, S.; BIENENSTOCK, E.; DOURSAT, R. Neural Networks and the Bias/Variance Dilemma. **Neural Computation**, MIT Press 238 Main St., Suite 500, Cambridge, MA 02142-1046 USA journals-info@mit.edu, v. 4, n. 1, p. 1–58, jan 1992. ISSN 0899-7667. Disponível em: <http://www.mitpressjournals.org/doi/10.1162/neco.1992.4.1.1>. Citado na página 35.

GESTEL, T. van.; BAESENS, B. **Credit risk management : basic concepts : financial risk components, rating analysis, models, economic and regulatory capital**. [S.l.]: Oxford University Press, 2009. 535 p. ISBN 9780199545117. Citado na página 52.

GONÇALVES, E. B.; GOUVÊA, M. A. COLLECTION SCORE E AS OPORTUNIDADES NO MERCADO DE NONPERFORMING LOANS. In: **Proceedings of the 12th CONTECSI International Conference on Information Systems and Technology Management**. TECSI, 2015. ISBN 978-859969310-0. Disponível em: <http://www.contecsi.fea.usp.br/envio/index.php/contecsi/12CONTECSI/paper/view/2263>. Citado na página 33.

GRIGORCHUK, T.; MAKSIMENKO, Z.; ROZANOVA, L.; BIKBULATOVA, G. SCORING MODELLING OF COLLECTION FINANCIAL FLOWS. **Oil and Gas Business**, n. 5, p. 630–655, oct 2015. ISSN 1813503X. Citado na página 33.

HANCOCK, M. F. **Practical Data Mining**. Taylor & Francis, 2011. ISBN 9781439868362. Disponível em: <https://books.google.com.br/books?id=syJLYgEACAAJ>. Citado nas páginas 39 e 41.

HE, H.; MA, Y. **Imbalanced learning : foundations, algorithms, and applications**. [S.l.: s.n.], 2013. ISBN 9781118074626. Citado na página 62.

Hoist Finance. **Amicable settlement - Hoist Finance**. 2017. Disponível em: <http://hoistfinance.com/about-hoist-finance/amicable-settlement/>. Citado na página 24.

KENNETH, C. L.; LAUDON, J. **Sistemas de Informações Gerenciais**. [S.l.]: São Paulo: Editora Pearson Brasil, Edição, 2007. Citado na página 37.

KUHN, M.; KUHN, M. The caret Package. 2009. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.150.2466>. Citado na página 59.

KUMAR, A.; BECK, T.; CAMPOS, C.; CHATTOPADHYAY, S. **Assessing Financial Access in Brazil**. The World Bank, 2005. Disponível em: <https://econpapers.repec.org/RePEc:wbk:wbpubs:7452>. Citado na página 29.

LABATUT, V.; CHERIFI, H. Evaluation of performance measures for classifiers comparison. **arXiv preprint arXiv:1112.4133**, 2011. Citado na página 62.

LARIVIERE, B.; POEL, D. V. den. Predicting customer retention and profitability by using random forests and regression forests techniques. **Expert Systems with Applications**, v. 29, n. 2, p. 472–484, 2005. ISSN 0957-4174. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0957417405000965>. Citado nas páginas 31 e 33.

LAWRENCE, D. B. **O Negócio de Crédito ao Consumidor - Risco e Recompensa.** [S.l.]: Citicorp, 1984. Citado nas páginas 30 e 31.

LIANG, D.; TSAI, C.-F.; WU, H.-T. The effect of feature selection on financial distress prediction. **Knowledge-Based Systems**, Elsevier, v. 73, p. 289–297, 2015. Citado na página 39.

LIAW, A.; WIENER, M. Classification and Regression by randomForest. **R News**, v. 2, n. 3, p. 18–22, 2002. Disponível em: <https://cran.r-project.org/doc/Rnews/>. Citado na página 59.

LIEW, L. **Optimizing Your Trading Strategy to 2 million in Profits! - What is**. 2016. Disponível em: <https://algotrading101.com/blog/1543426/what-is-curve-fitting-overfitting-in-trading-optimization>. Citado na página 36.

LU, H.; SETIONO, R.; LIU, H. NeuroRule: A Connectionist Approach to Data Mining. jan 2017. Disponível em: <http://arxiv.org/abs/1701.01358>. Citado na página 62.

MACHADO, A. R. Collection scoring via regressão logística e modelo de riscos proporcionais de Cox. 2016. Citado nas páginas 33, 34 e 48.

MASTERS, T. **Signal and Image Processing with Neural Networks: A C++ Sourcebook**. [s.n.], 1994. ISBN 9780471049630. Disponível em: <https://tywdphlhh.updog.co/dHl3ZHBobGhoMDQ3MTA0OTYzOA.pdfhttps://books.google.co.in/books?id=F8RQAAAAMAAJ>. Citado na página 57.

MEDEIROS, K. M. de; BRITO, F. I.; ARAUJO, A. O. Gestão de Crédito e Cobrança: análise dos resultados da terceirização em uma financeira. 2008. Citado na página 32.

MELLO, R. F. de; PONTI, M. A. **Machine Learning: A Practical Approach on the Statistical Learning Theory**. [S.l.: s.n.], 2018. XV, 362 p. ISSN 10450823. ISBN 9781577354260. Citado na página 57.

MITCHELL, T. M. **Decision Tree Learning**. 1997. 52–80 p. Citado na página 54.

OLIVEIRA, C. S.; SERPA, C. Reamostragem e imputação de dados em caso de eventos raros. dec 2013. Disponível em: <http://bdm.unb.br/handle/10483/8148?mode=full>. Citado na página 47.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, É. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, v. 12, n. Oct, p. 2825–2830, 2011. ISSN ISSN 1533-7928. Disponível em: <http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html>. Citado na página 45.

PEDRO, J. S.; PROSERPIO, D.; OLIVER, N. MobiScore: Towards Universal Credit Scoring from Mobile Phone Data. In: RICCI, F.; BONTCHEVA, K.; CONLAN, O.; LAWLESS, S. (Ed.). **User Modeling, Adaptation and Personalization**. Cham: Springer International Publishing, 2015. p. 195–207. ISBN 978-3-319-20267-9. Citado nas páginas 30 e 33.

PERRIER, A. **Effective Amazon Machine Learning.** [S.l.]: Packt Publishing, 2017. 298 p. ISBN 9781785881794. Citado na página 63.

Ponti Jr, M. P. Combining classifiers: from the creation of ensembles to the decision fusion. In: IEEE. **Graphics, Patterns and Images Tutorials (SIBGRAPI-T), 2011 24th SIBGRAPI Conference on**. [S.l.], 2011. p. 1–10. Citado na página 53.

PRATI, R. C.; BATISTA, G. E. A. P. A.; SILVA, D. F. Class imbalance revisited: a new experimental setup to assess the performance of treatment methods. **Knowledge and Information Systems**, Springer London, v. 45, n. 1, p. 247–270, oct 2015. ISSN 0219-1377. Disponível em: <http://link.springer.com/10.1007/s10115-014-0794-3>. Citado na página 57.

PROVOST, F.; FAWCETT, T. **What you need to know about data mining and Data Analytics**. [S.l.]: O'Reilly, 2013. 386 p. ISSN 0743-7463, 1520-5827. ISBN 9788578110796. Citado na página 65.

QI, Y. Random Forest for Bioinformatics. **Ensemble Machine Learning**, p. 307–323, 2012. Citado na página 54.

SHALEV-SHWARTZ, S.; OTHERS. Online learning and online convex optimization. **Foundations and Trends®in Machine Learning**, Now Publishers, Inc., v. 4, n. 2, p. 107–194, 2012. Citado na página 37.

Silva Santo, J. Análise de crédito e gestão do contas a receber na empresa TNT Mercúrio. 2013. Disponível em: <http://acad.saomarcos.br/rsm/bitstream/123456789/65/1/Jairo-Silva-dos-Santos.pdf>. Citado na página 32.

STRAUSS, R. E.; ATANASSOV, M. N.; De Oliveira, J. A. Evaluation of the principal-component and expectation-maximization methods for estimating missing data in morphometric studies. **Journal of Vertebrate Paleontology**, Taylor & Francis, v. 23, n. 2, p. 284–296, 2003. Citado na página 38.

SZCZERBA, M.; CIEMSKI, A. Credit Risk Handling in Telecommunication Sector. In: PERNER, P. (Ed.). **Advances in Data Mining. Applications and Theoretical Aspects**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. p. 117–130. ISBN 978-3-642-03067-3. Citado nas páginas 30 e 33.

THERNEAU, T.; ATKINSON, B.; RIPLEY, B. rpart: Recursive Partitioning. 2011. Citado na página 59.

Wei Fan; Janek Mathuria; Chang-Tien Lu. Making Data Mining Models Useful to Model Non-paying Customers of Exchange Carriers. In: **SIAM International Conference on Data Mining**. Newport Beach CA: [s.n.], 2005. p. 486–490. Citado na página 26.

WITTEN, I. H.; FRANK, E.; HALL, M. a. **Data Mining: Practical Machine Learning Tools and Techniques**. [S.l.: s.n.], 2011. 664 p. ISSN 14337851. ISBN 0080890369. Citado na página 53.

XIE, Y.; LI, X.; NGAI, E. W. T.; YING, W. Customer churn prediction using improved balanced random forests. **Expert Systems with Applications**, v. 36, n. 3, Part 1, p. 5445–5449, 2009. ISSN 0957-4174. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0957417408004326>. Citado nas páginas 31, 33, 34 e 48.

ZENG, S.; MELVILLE, P.; LANG, C. A.; BOIER-MARTIN, I.; MURPHY, C. Using Predictive Analysis to Improve Invoice-to-cash Collection. In: **Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. New York, NY, USA:

ACM, 2008. (KDD '08), p. 1043–1050. ISBN 978-1-60558-193-4. Disponível em: <http://doi.acm.org/10.1145/1401890.1402014>. Citado nas páginas 32 e 33.

ZHANG, Y.; KISIELIUS, V. **Method and system for providing a virtual collections call center system**. Google Patents, 2009. Disponível em: <https://www.google.com.br/patents/US20100010861http://www.google.tl/patents/US20100010861>. Citado na página 24.

GLOSSÁRIO

API: Application Programming Interface é uma interface de programação de aplicativos, cujo o papel principal é fornecer formas para integração em outros aplicativos, permitindo terceiros utilizarem os serviços da aplicação exteriormente..

ASR: Automatic Speech Recognition é uma tecnologia que permite um ser humano utilizar sua voz para comunicar com uma interface do computador, geralmente uma abordagem comumente utilizada é a conversão de fala para texto..

CRM: Customer Relationship Management é uma abordagem do qual utiliza-se um software para realizar Gestão do Relacionamento com o Cliente. Sendo assim o objetivo principal é colocar o cliente no centro dos processos da empresa de modo a viabilizar aquele tipo de percepção que permite antecipar as necessidades atuais e potenciais do cliente..

MIS: Management Information System em português Sistemas de Informação de Gestão, são sistemas e diretrizes responsáveis pelo apoio a tomada de decisão e gestão de uma empresa..

NPLS: Non-performing loans significa créditos não performados, ou não produtivos, são créditos cedidos por uma instituição financeira a um cliente e o mesmo depois de em médias 90 dias não realizou o pagamento das prestações acordadas devido uma dificuldade financeira, como a perda de um emprego por exemplo..


ICMC
USP
SÃO CARLOS