Revista de Sistemas de Informação da FSMA
n. 25 (2020) pp. 28-41

SISTEMAS DE INFORMAÇÃO
http://www.fsma.edu.br/si/sistemas.html

TRILHA PRINCIPAL ERI·GO

Avaliação de Crédito de Empréstimos Pessoais usando Técnicas de Aprendizado de Máquina

Pablo Simões Nascimento, *Pós-graduação em Ciência de Dados com Big Data*,
Karin Satie Komati, *Programa de Pós-Graduação em Computação Aplicada (PPComp)*,
Jefferson Oliveira Andrade, *Programa de Pós-Graduação em Computação Aplicada (PPComp)*,
Campus Serra do Instituto Federal do Espírito Santo (Ifes)

***Resumo*—Este trabalho consiste na proposta de uma arquitetura e de uma comparação de algoritmos de regressão para análise de propostas de empréstimo pessoal utilizando técnicas de aprendizado de máquina. O estudo de caso deste trabalho versa sobre os dados de uma financeira que possui uma base de dados histórica de análises de crédito, formada pelas informações de perfil do cliente, seu dados de relacionamento com a financeira e os dados da proposta de empréstimo. É importante que a arquitetura proposta atinja os seguintes objetivos: diminua o custo associado ao acesso aos dados de órgãos de proteção ao crédito, melhore a acurácia comparada ao processo existente, permita que a equipe de TI mantenha uma variável de controle sobre a aprovação ou não da proposta e que mantenha os empregos dos colaboradores. Para atender a todos estes requisitos, há duas etapas de análise, uma sem os dados de órgãos de proteção ao crédito e uma segunda com este tipo de dado. A divisão em duas etapas é um diferencial desta proposta, outro diferencial é que serão usados modelos de regressão cujos resultados são convertidos em resultados discretos via comparação com limiares de aprovação ou reprovação. Na primeira etapa, o resultado é "aprovado" quando estiver acima de um limiar de aprovação, "reprovado" quando estiver abaixo de um limiar de reprovação ou segue para a segunda etapa. O processo da segunda etapa é similar, mas os casos que não se enquadrem em aprovados ou reprovados seguem para análise manual. O trabalho compreende análise exploratória, pré-processamento dos dados, calibração, validação e testes em dois modelos de regressão: Floresta Aleatória e Mínimos Quadrados Parciais. Os resultados dos experimentos mostraram que a Floresta Aleatória apresentou resultados melhores que Mínimos Quadrados Parciais, e melhor que o sistema existente na financeira. Para esta configuração, a primeira etapa do processo de classificação é capaz de classificar 86,56% das propostas sem intervenção manual, e na segunda etapa mais 4,04% também são classificadas automaticamente, além de atingir 97% de acurácia ao final da segunda etapa.**

***Palavras-chave*—Análise de Crédito, Aprendizado de Máquina, Floresta Aleatória, Mínimos Quadrados Parciais.**

Autor correspondente: Jefferson Oliveira Andrade, *jefferson.andrade@ifes.edu.br*

Credit Assessment for Personal Loans using Machine Learning Techniques

***Abstract*—This paper describes the proposal of an archtecture and the comparison of regression methods for the analysis of personal loan proposals using machine learning algorithms. The case study presented here deals with data from a financial company that has a historical database of credit analyses, formed by the client's profile information, his previous relationship history with the financial company and the loan proposal data. It is important that the proposed architecture achieves the following objectives: decrease the cost associated with accessing credit protection agency's data, improve the accuracy compared to the existing process, allow the IT team keep a control variable over the approval or not of the proposal, and keep current employees' jobs. In order to meet all these requirements, a differential of this proposal is the usage of two stages of analysis, one without the data from credit protection agencies and the second with this type of data (if necessary). Another differential is that regression models will be used, whose results are converted into discrete results via comparison with denial/approval thresholds. In the first step, the result is *approved* – when it is above a pass threshold, *denied* – when it is below a fail threshold, or proceed to the second step. The second stage process is similar, except that cases that do not qualify as approved or denied continue for manual analysis. The methodology comprises exploratory analysis and pre-processing of data, calibration, validation and tests in two regression models: Random Forest and Partial Least Squares. The results of the experiments showed that the Random Forest achieved results that were better than both the Partial Least Squares and the existing system in the financial company. For this configuration, the first stage of the classification process is able to classify 86.56% of the proposals without manual intervention, and in the second stage, 4.04% are also classified automatically, in addition to reaching 97% accuracy at the end of the second stage.**

***Index Terms*—Credit Analysis, Machine Learning, Random Forest, Partial Least Squares.**

I. Introdução

O empréstimo pessoal é um serviço financeiro realizado entre duas partes: o mutuário, i.e., a pessoa

que pede o empréstimo e o credor, i.e., uma instituição financeira (tipicamente) que concede o empréstimo. No empréstimo pessoal, o mutuário recebe um valor, concedido pelo credor, que deve ser devolvido com juros em um prazo determinado. Neste artigo, as concedentes de crédito, de quaisquer tipos, serão denominadas por financeiras e as pessoas físicas que solicitam o empréstimos pessoais por mutuários ou clientes [1].

A concessão de crédito consiste em um mecanismo de injeção de dinheiro no mercado e um dos seus efeitos iniciais é o aquecimento da economia, uma vez que aumenta o poder de consumo das pessoas [2]. Tais recursos financeiros possibilitam que pessoas realizem seus projetos pessoais e melhorem sua qualidade de vida, através, por exemplo, da aquisição de bens de consumo ou, de serviços ou, de investimentos em estudos e viagens.

No Brasil, até há pouco tempo, apenas empresas de sociedade de crédito autorizadas pelo Banco Central, i.e., basicamente os bancos, poderiam fazer empréstimos. A legislação vem sendo atualizada e as *fintech* (do inglês: *FINance and TECHnology*), que são *startups* que trabalham para inovar e otimizar serviços do sistema financeiro, também passaram a estar autorizadas a prestar concessões de créditos sem a mediação de um banco [3]. Até mesmo uma pessoa física pode abrir uma ESC (Empresa Simples de Crédito) e conceder empréstimos [4].

Para uma pessoa física obter um empréstimo há duas formas mais comuns [5]: Crédito Pessoal (CP) e Crédito Direto ao Consumidor (CDC). O CP é a modalidade de crédito em que o cliente faz uma solicitação de crédito (pedido de empréstimo) em uma das filiais da financeira e sai do estabelecimento com dinheiro em espécie. O CDC é a modalidade de financiamento em que o cliente compra um produto em uma loja parceira (da financeira) e tem seu produto financiado pela financeira. Em ambos os casos, o desejo do cliente de obter dinheiro ou financiamento de um produto é convertido em uma proposta de crédito.

Neste tipo de transação financeira, quando os mutuários deixam de realizar o pagamento das parcelas, se tornam inadimplentes [6]. A inadimplência é a situação que se estabelece quando o mutuário não restitui totalmente o montante que lhe foi emprestado, de acordo com as regras contratuais previamente estabelecidas.

A partir do momento em que a proposta de crédito é feita, começa a fase de análise de crédito do cliente. O empréstimo pode ou não ser concedido conforme as políticas internas de análise de crédito. A análise de crédito tem como objetivo avaliar a possibilidade de conceder crédito, verificando a veracidade das informações fornecidas pelo cliente e suas condições de honrar os compromissos financeiros e não incorrer em inadimplência.

O estudo de caso deste trabalho versa sobre os dados de uma financeira que se situa na capital do Espírito Santo. Fundada em 1992, possui vasta experiência no mercado de crédito e, em 2012, chegou a mais de 200 mil clientes. Possui um quadro de mais de 100 colaboradores distribuídos entre matriz e suas 18 filiais.

Nesta empresa, a fase de análise de crédito é composta por duas etapas: na primeira etapa, chamada de análise automática, um sistema avalia regras internas previamente definidas (conforme a experiência de um gerente de crédito e de dados estatísticos históricos) e pode ter três resultados:

1) Proposta Aprovada – o crédito é concedido;
2) Proposta Reprovada – o crédito é negado; ou
3) Proposta Em Análise – quando é repassada para que um analista de crédito humano verifique mais detalhes que o ajudem na decisão final de aprovar ou reprovar a proposta.

Quando a situação recai neste último resultado, Proposta Em Análise, segue-se à segunda etapa, chamada de análise manual ou mesa de crédito. No cenário original da financeira, a primeira fase, de análise automática, respondia a 52% das propostas e os outros 48% são analisados manualmente.

Considerando-se as limitações inerentes à forma original de trabalho da empresa, i.e., uma solução de baixa tecnologia altamente dependente da análise manual das propostas, segue-se o questionamento: "Qual o efeito do uso de aprendizagem de máquina na concessão de crédito na operadora financeira em questão?"

Como resposta a este questionamento, buscou-se desenvolver um método para a análise automática de propostas de crédito, que torne esta análise mais assertiva, deixando o processo mais rápido e eficiente, diminuindo o volume de propostas repassados para a mesa (análise manual) e consequentemente, diminuindo a demanda de todo o setor de crédito.

Isto posto, este trabalho se propõe a desenvolver uma comparação entre dois modelos baseados em técnicas de aprendizado de máquina que atendam aos requisitos acima, e ainda avaliar experimentalmente o desempenho destes modelos de aprendizado de máquina na concessão de crédito na operadora financeira em questão. O uso de técnicas de aprendizado de máquina neste tipo de aplicação é comum ao menos desde os anos de 1990 [7].

Entretanto, de forma diferente da usual, este trabalho apresenta um estudo de modelos de regressão em duas fases, cujos resultados são categorizados por limiares preestabelecidos. Esta forma de trabalho foi utilizada a pedido da empresa do estudo de caso. Além disso, ficou estabelecido que os limiares sejam reavaliados continuamente pela equipe de TI em conjunto com os analistas de crédito. Os modelos de regressão selecionados foram a Floresta Aleatória (do inglês *Random Forest*) [8] e Mínimos Quadrados Parciais (do inglês *Partial Least Squares*) [9].

Um importante diferencial desta proposta é o fato de que o método proposto é composto dois modelos de regressão. O primeiro modelo, tal como no sistema original, categoriza a proposta em Aprovada, Reprovada ou Em Análise; o segundo modelo avalia as propostas na situação Em Análise, levando em conta dados externos de consulta de crédito. Após a aplicação do segundo modelo ainda é possível que algumas propostas sigam para a análise manual. É um resultado importante que o sistema melhore a acurácia da avaliação das propostas de empréstimo, mas

também é importante que os empregos dos colaboradores da empresa sejam preservados. A acurácia é a soma dos positivos verdadeiros e negativos verdadeiros, dividido pelo total de propostas.

A proposta engloba técnicas de pré-processamento e limpeza dos dados, algoritmos de aprendizado de máquina para desenvolvimento dos modelos, bem como ferramentas de visualização de dados e verificação dos resultados para conclusões e validação dos modelos encontrados. Para os experimentos, utilizou-se uma base de dados interna da empresa contendo os perfis dos clientes, o histórico de propostas analisadas e informações sobre o cliente em órgãos de crédito.

Este artigo segue com a Seção II que descreve trabalhos relacionados; a Seção III explica a arquitetura proposta de forma mais detalhada; a Seção IV apresenta a base de dados inicial e as tarefas de análise exploratória e pré-processamento dos dados; a Seção V explica os modelos de regressão utilizados, bem como o processo de calibração, treinamento e validação; a Seção VI expõe os resultados dos experimentos, discutindo-os e, por fim, a Seção VII finaliza com as conclusões deste trabalho e algumas indicações de trabalhos futuros.

II. Trabalhos Relacionados

MUITOS trabalhos relacionados à procura por uma melhor proposta para a análise de crédito pessoal têm sido desenvolvidos nas últimas décadas.

Vasconcellos [10] aborda uma proposta para aperfeiçoamento da análise de concessões de crédito a pessoas físicas com o desenvolvimento de um modelo de análise preditiva. O modelo proposto por Vasconcellos parte de uma base de dados de uma instituição financeira em que o autor trabalha. Este trabalho utilizou o critério da inadimplência, que analisa o comportamento de pagamentos em empréstimos anteriores para definir a qualidade da decisão de concessão de crédito. Créditos considerados ruins eram os que apresentaram atraso de 61 ou mais dias no pagamento da prestação, enquanto os bons eram aqueles com atraso de no máximo 60 dias. Vasconcellos difere deste trabalho principalmente com relação ao critério utilizado para classificação da proposta de crédito que foi a inadimplência. Enquanto este trabalho propõe uma abordagem de classificação alternativa à utilizada atualmente e baseada em um histórico de propostas já classificadas, Vasconcellos propôs um modelo, também baseado no histórico de propostas recentes, porém, focando na qualidade dos créditos concedidos cujo resultado apresenta um indicador de tendência à inadimplência. O trabalho não se preocupou, portanto, em replicar o comportamento da atual análise de crédito feita pela instituição. Ao invés disso, buscou substituir tal modelo oferecendo um que seja capaz de classificar melhor a proposta com vistas a evitar alto índice de inadimplência. Faz a ressalva, contudo, que um acompanhamento a longo prazo se faz necessário devido ao risco quanto às possibilidades de redução de lucro com os créditos. Concluiu ainda que, embora a taxa de inadimplência seja reduzida como resultado de uma boa análise, reconhece que pode reduzir também o lucro que hoje é obtido com os pagamentos dos juros de quem esteve inadimplente em algum momento do contrato. Se reduz o número de inadimplentes, pode-se reduzir o lucro dos juros de inadimplência.

Em meados dos anos 2000 começou a aparecer uma nova modalidade de serviço financeiro denominados "empréstimos *peer-to-peer*" (P2P) ou "empréstimos ponto a ponto", que é a prática de emprestar dinheiro a indivíduos ou empresas por meio de serviços *online* que combinam credores com mutuários [11]. Os empréstimos *peer-to-peer* não são adequadamente classificados como nenhum dos três tipos tradicionais de instituições financeiras, i.e., captadores de depósitos, investidores e seguradoras, e às vezes são classificados como um *serviço financeiro alternativo*. Embora não sejam idênticos ao empréstimo pessoal, há similaridades entre os empréstimos *peer-to-peer* e o tipo de agente financeiro que foi alvo deste estudo.

Rodrigues, Brasil, Costa et al. [12] também investigaram técnicas de aprendizado de máquina para a análise de crédito no contexto dos empréstimos *peer-to-peer*. Objetivando a criação de melhores ferramentas de análise de risco no mercado de crédito, foi feita uma análise comparativa entre vários modelos de algoritmos de classificação utilizando dados fornecidos por uma plataforma de empréstimos *online Peer-to-Peer* chamada Lending Club. O trabalho selecionou os 9 atributos com maior peso sobre o resultado de predição de inadimplência, associando dados do mutuário com os de *credit scoring* na base de dados analisada e obteve resultados melhores do que os apresentados nos trabalhos correlatos. O trabalho difere deste em dois pontos principais, o primeiro ao não buscar treinar o modelo conforme as classificações de propostas de crédito já classificadas no histórico, mas buscar prover um modelo capaz de oferecer índice para inadimplência conforme os atributos analisados. O segundo ponto é que aquele trabalho buscou desenvolver análises comparativas dentre vários algoritmos de classificação buscando evidenciar qual melhor se aplica à predição de inadimplência, e este trabalho analisa algoritmos de regressão.

Polena e Regner [13] estudaram os determinantes da inadimplência dos mutuários nos empréstimos P2P, também utilizando um conjunto de dados de empréstimos do Lending Club. Em seu trabalho foram definidas quatro classes de risco de empréstimo e foi utilizada regressão linear para testar a significância das variáveis determinantes em cada classe de risco de empréstimo. Os resultados sugerem que o significado da maioria das variáveis depende da classe de risco do empréstimo. Apenas algumas variáveis são consistentemente significativas em todas as classes de risco. A relação dívida/renda, e consultas de crédito nos últimos 6 meses, por exemplo, estão correlacionados positivamente com a taxa de inadimplência, enquanto renda anual está correlacionada negativamente, independentemente do nível de risco.

Zhang, Li, Hai et al. [14] realizaram um estudo empírico usando o conjunto de dados público da Paipaidai, a maior operadora de empréstimo P2P *online* na China. Foi

utilizada regressão logística para analisar os fatores que determinam a probabilidade de obtenção do empréstimo. Seu resultado indica que a taxa de juros anual, o período de pagamento, a descrição, o grau de crédito, o número de empréstimos anteriores bem-sucedidos, e o número de empréstimos anteriores com falha são fatores significativos para o financiamento bem-sucedido na plataforma Paipaidai.

Wang, Ma, Huang et al. [15] propuseram dois diferentes métodos *ensemble*, i.e., métodos que utilizam múltiplos classificadores fazendo a fusão de seus resultados, baseados em árvores de decisão sobre subespaços randômicos, que são, em essência, variações do método tradicional de floresta aleatória utilizado no presente trabalho. Estes autores também demonstraram que, em seu trabalho, todas as variações de florestas aleatórias obtiveram melhor performance do que os cinco métodos de classificadores simples escolhidos para comparação: regressão logística, análise por discriminantes lineares, perceptron multi-camadas, e árvores de decisão.

No contexto de análise de crédito, conjuntos de dados desequilibrados ocorrem com frequência, pois o número de empréstimos inadimplentes em uma carteira é geralmente muito menor que o número de observações que são adimplentes. Brown e Mues [16] realizaram um estudo empírico sobre a sensibilidade de diversos classificadores ao desbalanceamento dos conjuntos de dados. Os resultados deste estudo indicam que os classificadores de floresta aleatória e de gradiente descendente têm um desempenho muito bom e são capazes de lidar comparativamente bem com os desequilíbrios acentuados de classes nos conjuntos de dados estudados. Os experimentos usaram as propostas clássicas, sem modificações, de floresta aleatória e de gradiente descendente.

Louzada, Ara e Fernandes [7] realizaram uma revisão sistemática da literatura sobre análise de crédito (*credit scoring*) abrangendo o período de 1992 à 2015. Foram analisados 187 artigos científicos sobre o tema, que foram categorizados de acordo com diversas dimensões, incluindo o propósito do artigo (proposição de novos métodos de análise de crédito, estudos de medição de performance, seleção de características, discussão conceitual, etc.) e o principal método de classificação utilizado no estudo (máquinas de vetor de suporte, redes neurais artificiais, regressão linear, árvores de decisão, métodos híbridos, métodos *ensemble*, etc.). Foi levantado que $51,3\%$ dos artigos propunham novos métodos de análise de crédito, enquanto algo em torno de 20% dos trabalhos tratavam de análises comparativas entre diferentes técnicas. Dentre as técnicas de *scoring*, o mais comum foi a utilização de métodos híbridos (que combinem 2 ou mais técnicas) com $19,8\%$, seguidos pelos trabalhos que propunham métodos *ensemble* com $14,6\%$. As máquinas de vetor de suporte ficaram em terceiro lugar com $13,5\%$, seguidas pelos trabalhos que propõem o uso de redes neurais artificiais com $12,5\%$. Talvez o resultado mais surpreendente do trabalho de Louzada, Ara e Fernandes [7] seja a constatação de que ao longo do período analisado, independentemente da técnica utilizada, não houve grandes variações na performance dos classificadores para a análise de crédito em situações gerais. O que parece sugerir que os ganhos eventuais obtidos são específicos do cenário de aplicações ou específicos dos agentes financeiros.

Todos os trabalhos citados usam base de dados com atributos diferentes dos encontrados na base de dados deste trabalho, e portanto os resultados das técnicas não podem ser transpostos ou comparados com esta proposta.

III. Arquitetura do Sistema Proposto

UM modelo preditivo supervisionado pode ser descrito como uma função que, a partir de um determinado conjunto de dados rotulados, constrói um estimador. O estimador pode ser definido como classificador ou regressor, dependendo do tipo de saída da função. Uma saída discreta caracteriza um classificador, e uma saída contínua caracteriza um regressor [17]. A natureza do problema discutido neste trabalho é de classificação, uma vez que cada proposta tem resultado discretos, sendo Aprovada, Reprovada ou Em Análise. No entanto, a empresa tomada como estudo de caso manifestou desejo de ter um índice que permitisse os ajustes dos limiares de aprovação e reprovação. Nesse cenário, aplica-se o estimador do tipo regressor que permite definir limites acima e abaixo dos quais classificar-se-ão as propostas e as que não forem classificadas continuarão demandando o setor de crédito com análise manual.

O processo de execução deste trabalho é ilustrada na Figura 1, já considerando os modelos treinados. Uma nova proposta de crédito será a entrada para o **Modelo 1**, que retornará um valor entre 0 e 1 (normalizado), quanto mais próximo de 1, maior a confiança de aprovação. Esse valor representa a confiança de uma proposta ser Aprovada (isto é, acima de um limiar de aprovação) ou de ser Reprovada (isto é, abaixo de um limiar de reprovação). Faz-se necessário um estudo dos limiares de aprovação e reprovação.

Chamamos de **Limiar** o valor escolhido entre 0 e 1 resultante da predição do modelo que estabelecerá um limite para classificação das propostas. Tal limite, uma vez escolhido, define o ponto de corte para aprovar ou reprovar as propostas conforme a regra abaixo:

$$\text{Proposta} = \begin{cases} \textbf{aprovada} & \text{se predição} \geq \text{limiar} \\ \textbf{reprovada} & \text{se predição} \leq (1 - \text{limiar}) \end{cases}$$

Portanto, o *limiar de aprovação* se refere ao limite superior acima do qual as propostas são aprovadas e o *limiar de reprovação* é o limite inferior equivalente $(1 - \text{limiar de aprovação})$ abaixo do qual as propostas são reprovadas.

No entanto, há propostas que podem estar entre estes limiares, e seguirão para uma segunda fase. Passa-se à consulta aos órgãos de proteção ao crédito, ou seja, consulta a entes externos. Órgãos de proteção ao crédito oferecem serviços de informações de crédito, disponibilizando dados de adimplência e inadimplência de pessoas físicas ou jurídicas. Dentro os órgãos mais conhecidos estão o Serviço

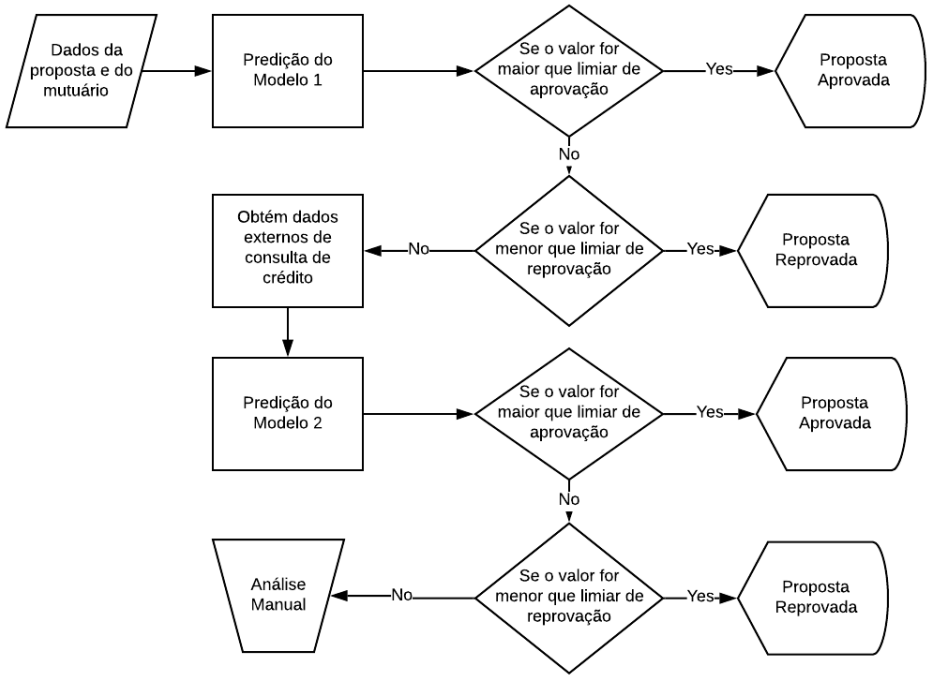


Fig. 1. Fluxo do processo proposto para análise das propostas de crédito.

de Proteção ao Crédito (SPC) e a Serasa [18]. Nem todos os clientes possuem informações disponíveis nos órgãos de proteção ao crédito, e existe um custo financeiro para cada consulta realizada.

A próxima etapa utiliza um novo modelo, o qual denominaremos de **Modelo 2**. Esta arquitetura permite consultar a situação do cliente junto aos serviços de informações de crédito apenas quando necessário, ou seja, apenas nos casos em que as propostas não tenham uma confiança suficiente para defini-las como Aprovadas ou Reprovadas, diminuindo o gasto financeiro na busca de tal informação.

Da mesma forma que no **Modelo 1**, o **Modelo 2** retornará um valor entre 0 e 1 (normalizado), e da mesma forma, quanto mais próximo de 1, maior a confiança de aprovação. Acima de um limiar de aprovação, será Aprovado e abaixo de um limiar de reprovação, será Reprovado. Propostas com valores entre os limiares seguem para análise manual.

Cabe notar que no desenvolvimento dos modelos utilizados na arquitetura acima, assumiu-se que a eventual ocorrência de variáveis endógenas[1] não acarretaria problemas relevantes. Em parte, esta suposição se justifica pelos métodos de aprendizado de máquina utilizados (floresta aleatória e quadrados mínimos parciais) serem descritos na literatura como robustos na presença de correlações entre as variáveis do modelo. Existe também evidência que sugere que o método de floresta aleatória apresenta bons resultados, mesmo na presença de endogeneidade [19], [20].

[1]Uma variável exógena refere-se a uma variável que é determinada fora do modelo e representa as entradas de um modelo. Em contraste, variáveis endógenas são determinadas dentro do modelo e, portanto, representam as saídas de um modelo.

IV. Conjunto de dados

O conjunto de dados inicial é formada por quase 200.000 propostas, resultado da seleção de registros de 14 meses de operação da financeira, entre janeiro de 2018 e março de 2019. O conjunto de dados é composto por 47 atributos distribuídos conforme mostrado na Tabela I. O atributos são divididos em 3 tipos: referente ao perfil do cliente, referente ao histórico de relacionamento e dados da proposta.

Como a Tabela I mostra, no perfil do cliente há 3 subgrupos, relacionados ao indivíduo, à residência e à renda. Além dos dados comuns do indivíduo (sexo, idade e situação conjugal), há o atributo "Código do cliente" que não tem valor semântico, e é gerado automaticamente pelo sistema gerenciador de banco de dados. O atributo "Tipo do Cliente" indica se é a primeira proposta do cliente ou se ele já teve proposta aprovada anteriormente. Quanto ao atributo "CPF divergente da UF de residência", se deve ao fato de que o antepenúltimo dígito do CPF indica o estado em que o CPF foi emitido:

0) Rio Grande do Sul
1) Distrito Federal, Goiás, Mato Grosso, Mato Grosso do Sul e Tocantins
2) Amazonas, Pará, Roraima, Amapá, Acre e Rondônia
3) Ceará, Maranhão e Piauí
4) Paraíba, Pernambuco, Alagoas e Rio Grande do Norte
5) Bahia e Sergipe
6) Minas Gerais
7) Rio de Janeiro e Espírito Santo

TABELA I
LISTAGEM DOS ATRIBUTOS, ORGANIZADOS POR GRUPOS DE DADOS.

Informação	Atributos
15 atributos cadastrais do cliente	**Individuais**: Código do cliente, Idade, Sexo, Situação conjugal, Tipo de cliente (Novo ou Recompra), CPF divergente da UF de residência
	Residenciais: Localidade, UF de residência, Reside em UF diferente da filial
	Renda: Valor do salário, Valor de outras rendas, Valor de salário do cônjuge, Valor total da renda, Tempo de serviço, Tipo de atividade
17 atributos sobre o histórico de relacionamento	**Dados quantitativos**: Quantidade de acordos de cobrança em aberto, Quantidade de contratos em fraude, Quantidade contratos não quitados, Quantidade de propostas anteriores aprovadas, Quantidade contratos quitados antes da proposta, Quantidade contratos antes da proposta, Quantidade de propostas analisadas recentemente, Quantidade de acordos
	Dados de atrasos: Fiador em atraso, Parcelas atrasadas entre 15 e 30 dias, Parcelas atrasadas a mais de 30 dias
	Percentuais de quitação: Percentual quitado do primeiro contrato, Percentual quitado último contrato
	Dados gerais: Fora da praça de atuação, Possui pendência jurídica, Margem bruta, Cliente sem contrato em aberto
15 atributos sobre a proposta	Data e hora da proposta, Código do estabelecimento, Tipo de operação (CP ou CDC), Forma de pagamento, Valor à vista, Valor de entrada, Valor a financiar, Quantidade de parcelas, Valor total da proposta, Valor em tributos, Percentual mensal, Percentual de juros mensal, Situação da proposta, Situação da biometria, Condição comercial

8) São Paulo
9) Paraná e Santa Catarina

Um aspecto a ser considerado sobre os dados é a sua temporalidade. Alguns dados cadastrais dos clientes são atualizados apenas quando uma nova proposta de crédito é solicitada, e.g., o valor do salário ou renda. Esse valor é atualizado no banco de dados e não se mantém o histórico dessa alteração. Portanto, o valor da renda de um cliente informado numa proposta em 2018 pode não mais ser o mesmo valor em 2019 quando fizer uma nova proposta, pois o valor pode ter sido atualizado. Para contornar essa limitação, uma condição foi necessária no filtro de todas as propostas: apenas as propostas de clientes que não tiveram cadastro atualizado foram selecionadas. Assim, garante-se que o resultado de avaliação daquela proposta foi obtido considerando exatamente as mesmas informações que o modelo atual utilizou na época em que a proposta foi gerada.

O conjunto de atributos sobre o histórico do relacionamento é subdividido em: dados quantitativos de propostas, contratos e acordos; dados sobre atrasos; percentuais de quitação e dados gerais. No grupo de dados da proposta, consta o "Código do Estabelecimento" que são códigos da empresa parceira. A financeira tem parceria com vários estabelecimentos comerciais em diversas áreas, tais como eletrodomésticos e móveis. Quando um cliente opta pela compra de um móvel na empresa parceira, por exemplo, pode optar por fazer o CDC via financeira. A diferença entre "Percentual mensal" e "Percentual de juros mensal" é que a primeira representa o custo efetivo total, que é a taxa percentual que demonstra o somatório de todos os custos que o cliente terá em uma operação de credito; já o segundo é o percentual de juros mensal que incide sobre o valor a vista que resulta no valor financiado final.

Em algumas empresas parceiras e nas filiais da financeira é possível fazer a captura da face do cliente. Há um sistema separado que faz o reconhecimento facial. Esta informação é armazenada no atributo "Situação da Biometria", que pode ter os seguinte valores "(SI) SEM IMAGEM" (face não foi coletada), "(CA) COM ALERTA" (cliente com alerta na análise de biometria facial, reconhecido como cliente com suspeita de fraude) e "(SA) SEM ALERTA" (cliente reconhecido, mas sem suspeita de fraude).

Na base de dados da financeira as propostas possuem as seguintes situações, conforme já citado: Aprovada, Reprovada e Em Análise. Além destas, a proposta pode ainda estar *Cancelada* ou *Pré-Aprovada*. Uma proposta é classificada como *Cancelada* quando o cliente iniciou o processo de solicitação de crédito, mas não avançou na entrada de dados e mesmo após contato telefônico não quis continuar com o processo, e assim, não se tem um cadastro completo da pessoa. Uma proposta Em Análise ainda está em análise manual e, portanto, ainda não há uma avaliação se o crédito será ou não concedido. A proposta recebe o status de *Pré-Aprovado* quando há uma campanha de marketing para o cadastramento de pessoas como futuros clientes da financeira, mas que ainda não efetivaram concessão de crédito.

Para o propósito deste trabalho apenas propostas Aprovadas ou Reprovadas foram selecionadas. A base de dados é razoavelmente bem balanceada possuindo 62% das propostas reprovadas e 38% aprovadas.

No processo original da financeira, aproximadamente 40 regras condicionais são avaliadas envolvendo cadastro do cliente, histórico de contratos junto a empresa e *score* de crédito na praça. Os resultados de algumas regras reprovam imediatamente, outras aprovam imediatamente, outras ainda repassam para a mesa de crédito imediatamente sem necessidade de passar por todas as regras. A regra pode ser simples envolvendo apenas uma característica (por exemplo, se a idade do cliente é maior que

18 anos) ou pode ser complexa envolvendo cálculos com várias características.

As regras do processo original da financeira são executadas por um sistema independente ao qual é submetida a proposta do cliente. Infelizmente, não é gravado um registro do resultado de todas as regras para cada proposta submetida, apenas o resultado final, que aprovou, reprovou ou repassou a proposta. Diante desse cenário no qual as regras não puderam ser reconstituídas completamente, nem os resultados das avaliações das regras puderam ser aproveitados, foi necessário gerar uma base de dados independente, mas que representasse com máxima proximidade as informações que foram levadas em consideração na época em que as propostas foram analisadas.

A. Análise exploratória e Pré-processamento dos dados

De acordo com Batista [21], considera-se que a análise exploratória de dados é um processo semiautomático, isto é, depende da capacidade da pessoa que a conduz em identificar os problemas presentes nos dados, e utilizar os métodos mais apropriados para solucionar cada um dos problemas.

A análise exploratória de dados deste trabalho foi realizada utilizando a linguagem Python 3, com pacotes Pandas, Scikit Learn e NumPy, na plataforma do Jupyter Notebook.

Alguns dados precisaram ser tratados, como a remoção de *outliers*, tratamento dos valores nulos e enriquecimento de semântica pela criação derivada de novos dados. A base de dados inicial contém várias inconsistências, tais como registros com quase todos os valores de atributos como nulos e cadastro de clientes menores que 18 anos. Estes registros foram retirados. Além disso, ao se verificar a presença de valores muito discrepantes, foi realizada a remoção de *outliers*. Havia valores para os atributos "valor a financiar" e "valor da renda" extremamente atípicos, e as respectivas observações (registros) foram excluídos da base de dados.

A extração de dados inicial foi realizada com a data da proposta, que tem o dia e a hora em que o cliente fez a solicitação. Após alguns testes preliminares, constatou-se que os modelos exploratórios apresentavam melhor acurácia com a quebra da data em campos de informação mais granulares: o dia, o mês, o ano e a hora. Esse ajuste resultou em uma melhora de acurácia de 2% comparado ao atributo único. Um achado inesperado foi a constatação de que propostas feitas pela manhã têm maior chance de serem aprovadas que aquelas feitas à tarde, tendência que foi confirmada posteriormente pelos analistas da financeira.

Foram verificadas as frequências de todos os atributos de tipos categóricos, e identificados aqueles com frequências muito baixas (indicando, provavelmente, erros de digitação). Os atributos categóricos com número de ocorrências bem distribuídas foram mantidos, mas foram convertidos em valores numéricos utilizando a função *LabelEncoder*. Esta função converte os valores categóricos em valores inteiros correspondentes, assim, a identificação do tipo de empréstimo CP ou CDC seria alterada para 0 e 1, por exemplo.

B. Seleção de Atributos

O conjunto de atributos fornecido e utilizado inicialmente para avaliação foi exatamente o mesmo que o utilizado pelo processo original de análise automática da financeira. Sendo assim, nossa suposição é de que a construção de um modelo preditivo cujo aprendizado seja dirigido pelas mesmas informações da análise original, e que use como base de comparação os resultados daquela análise, apresente respostas comparativamente semelhantes.

A seleção de atributos foi feita com base no coeficiente de correlação de Pearson [22]. Foram mantidos apenas os atributos cujo coeficiente de correlação fosse menor que 95%, do contrário julga-se que as informações são redundantes. Para os atributos redundantes, ou seja, com correlação maior ou igual a 95% apenas um atributo foi selecionado, de forma aleatória, e os outros excluídos da base de dados. Desta forma, 5 atributos foram excluídos por esta seleção:

1) Quantidade de contratos antes da proposta,
2) Valor a financiar,
3) Valor total da proposta,
4) Valor em tributos e
5) Percentual de juros mensal.

Outro critério de análise empregado foi a variedade de ocorrências de cada valor para todos atributos restantes. Identificou-se as características sem variabilidade, onde quase todos registros possuíssem o mesmo valor. O atributo "Forma de Pagamento", por exemplo, possuía o valor "Boleto" na maioria dos registros (99,3%) e o valor "Cheque" em 0,7% dos registros. Segundo esse critério, 6 características foram excluídas da base de dados:

1) Forma de Pagamento,
2) Fora da praça de atuação,
3) Fiador em atraso,
4) Quantidade de acordos de cobrança em aberto,
5) Quantidade de contratos em fraude e
6) Possui Pendência Jurídica.

Ao final, o conjunto de dados **Base de Dados Pré-processada** de experimentos foi formado por 39 atributos e 192.177 registros. Dos 47 atributos iniciais foram retirados 11 através da seleção de atributos, mas a transformação da data para atributos separados para suas partes gerou 3 atributos a mais.

C. Dados de órgãos de proteção ao crédito

As informações de crédito externas são fornecidas por órgãos de crédito como SPC, Boa Vista ou Serasa, pois a financeira do estudo de caso mantém contrato de acesso à dados com estes 3 órgãos.

Existem dois tipos de informações principais fornecidas: de *classificação* (neste trabalho, usou-se o nome deste atributo em itálico para não se confundir com a tarefa de classificação de um modelo preditivo) e de *scoring* de crédito do cliente na praça. *Classificação* é a situação de crédito do cliente que informa se está:

- "negativado",
- "normal" (mas com passagem, já tendo sido negativado anteriormente),
- "com algum alerta" e,
- "nada consta".

Scoring é um valor numérico de 0 a 1.000 que atribui uma pontuação ao cliente quanto ao seu risco de inadimplência em uma eventual concessão de crédito. Tal pontuação é específica para cada órgão, ou seja, 500 pontos no SPC não significam o mesmo risco que 500 pontos no Serasa. Além disso, a região também influencia no quão boa uma pontuação é: por exemplo, 500 pontos no Espírito Santo pode ser uma boa pontuação, mas não necessariamente o será em outro estado.

Neste trabalho, foi utilizada apenas a informação de *classificação* e em qual órgão foi feita a consulta: SPC, Boa Vista ou Serasa. Assim, para o **Modelo 2**, foram incluídos os atributos *Órgão da Classificação* e *Classificação*.

V. Modelos de Regressão

NESTA seção apresentam-se dois modelos de regressão: Floresta Aleatória e Mínimos Quadrados Parciais. Além disso, descrevem-se o processo de calibração, treinamento e validação. A literatura descreve diferentes formas, utilizando diferentes percentuais, para a divisão do conjunto de dados para cada um dos processos de calibração, treinamento e validação [23], [24], e não há uma regra única padrão. Para este trabalho, a divisão da base de dados é ilustrada na Figura 2. Na figura, a largura dos retângulos de cantos arredondados representam o tamanho da base de dados.

A Seção IV descreveu o conjunto de dados inicial e o processo de ETL (*Extract-Transform-Load*) utilizado para a obtenção do conjunto de dados que foi utilizado neste trabalho, que ficou com tamanho um pouco menor que o conjunto inicial. O conjunto de dados foi dividido em duas partes: a primeira para treino e validação do modelo (denominada de **Treino Inicial**) com 96.136 registros e a segunda para os experimentos (denominada de **Testes**) com 96.041 registros, uma divisão de aproximadamente 50–50% [25], [26], [27]. Esta proporção é considerada como "ambiciosa" por Aguilera, Guardiola-Albert e Serrano-Hidalgo [28], pois comumente se usam proporções maiores para treinamento e menores para testes, o qual tende a melhorar a acurácia final da classificação.

A base **Testes** não é usada em nenhum momento durante o processo de calibração, treino e validação. Este subconjunto é usado somente para os experimentos finais, já com os modelos calibrados, treinados e validados.

Para o modelo de Floresta Aleatória usou-se como método de validação a técnica de *holdout*, dividiu-se a base de **Treino Inicial** em duas partes: 70% para treino (base **Treino**) e os outros 30% para validação (base **Validação**).

A. Regressão via Floresta Aleatória

Floresta Aleatória (do inglês *Random Forest*) é um algoritmo de aprendizagem supervisionado [8], [29]. É um método *Ensemble* do tipo *Bagging*, isto é, que constrói vários modelos (árvores de decisão) em paralelo, a partir de diferentes subamostras do conjunto de dados de treinamento. Assim, a Floresta Aleatória consiste em um grande número de árvores de decisão individuais que funcionam como um conjunto. Cada árvore individual na floresta aleatória apresenta uma previsão e o resultado com mais votos torna-se a previsão final do modelo.

Uma grande vantagem da combinação entre diversas árvores de decisão é a redução dos erros que árvores de decisões individuais podem obter, devido à sua sensibilidade a ruídos. Enquanto algumas árvores podem estar erradas, muitas outras árvores estarão certas, então, como um grupo, as árvores podem se mover na direção correta.

O método de Floresta Aleatória pode ser utilizado tanto para regressão quanto para classificação. Para o caso de regressão, uma especialização deste algoritmo é necessária. Chamado de *Random Forest Regressor*, este método fornece como saída um valor contínuo de predição.

1) Calibração da Floresta Aleatória: O algoritmo de treino do modelo *Random Forest Regressor* é sensível a dados não normalizados, logo, é necessário normalizá-los. Para esta tarefa a classe *StandardScaler* da biblioteca *Scikit Learn* foi utilizada. O escalonador padrão (*StandardScaler*) do *Scikit Learn* transforma os dados para média próxima de zero e um desvio padrão próximo de um, assumindo que não há valores discrepantes nos dados normalizados, o que é verdade no nosso caso pois realizamos o processo de extração de outliers na fase de ETL.

Para construção do modelo regressor de floresta aleatória foi usada a classe **RandomForestRegressor** do módulo **ensemble** da biblioteca *Scikit Learn*. Este modelo é treinado passando-se três parâmetros: o número de árvores, os dados de treino e o vetor de saída. A fim de encontrar qual a quantidade de árvores que nos forneça o "melhor" modelo, foi realizada uma calibração ou *tunning* [12]. O gráfico da Figura 3 mostra o resultado do *score* (resultado da função de calibração) pela quantidade de árvores, quanto mais perto de 1, melhor o resultado. A partir de 100 árvores, o *score* se mantém praticamente constante. Assim, foi gerado um modelo 100 árvores no parâmetro, com *score* de 0,77.

A mesma metodologia foi usada na calibração do **Modelo 2**, com a diferença de que este modelo conta com os atributos de informações extras sobre a *classificação* do cliente em um órgão de proteção ao crédito. Somente propostas com a informação de *classificação* foram utilizadas para o treinamento deste modelo. Do contrário, estaríamos treinando o modelo com os mesmos dados do modelo anterior, o que geraria um modelo incorreto. Assim, para o **Modelo 2**, filtrou-se a base de dados com registros que possuíssem uma consulta. O gráfico da Figura 4 mostra que a partir de 60 árvores o crescimento do *score* não melhora muito o modelo, assim, o **Modelo 2** ficou com 60 árvores.

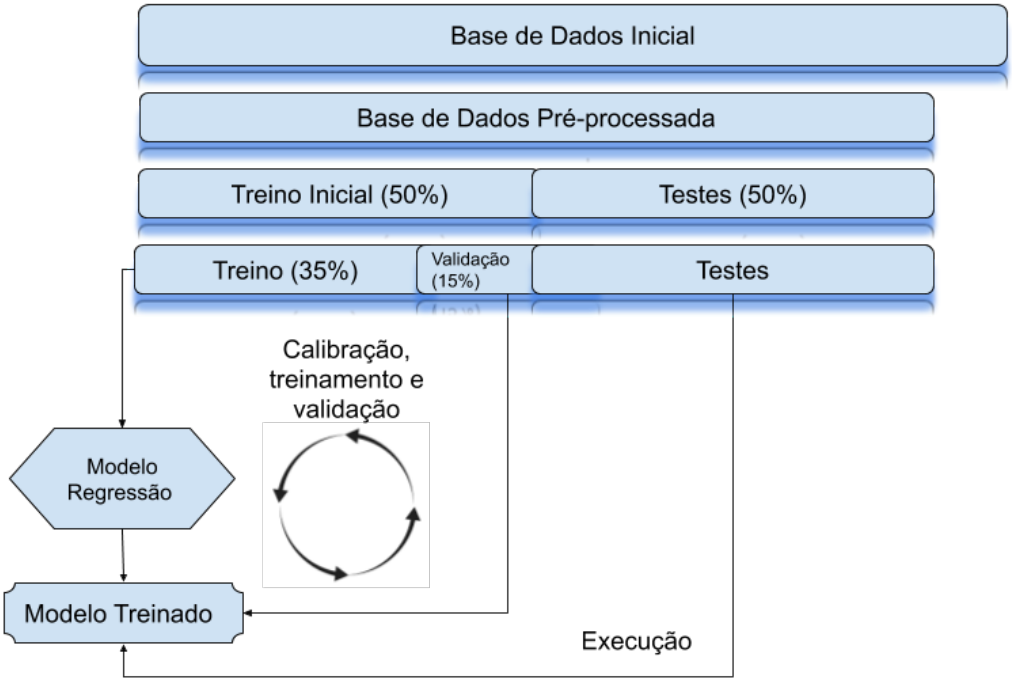


Fig. 2. Metodologia da calibração, treinamento, validação e testes.

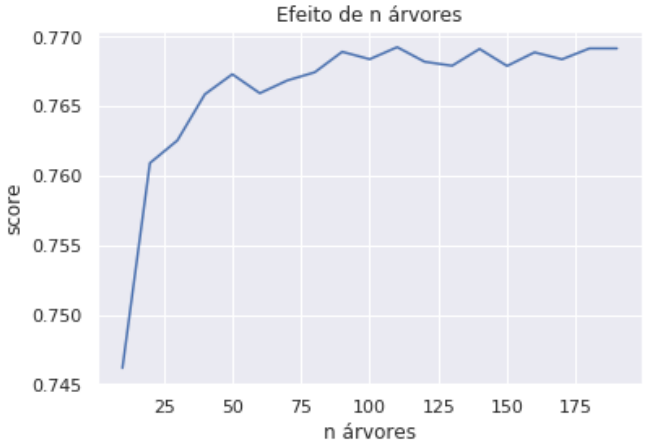


Fig. 3. Gráfico do *score* pelo número de árvores do **Modelo 1**.

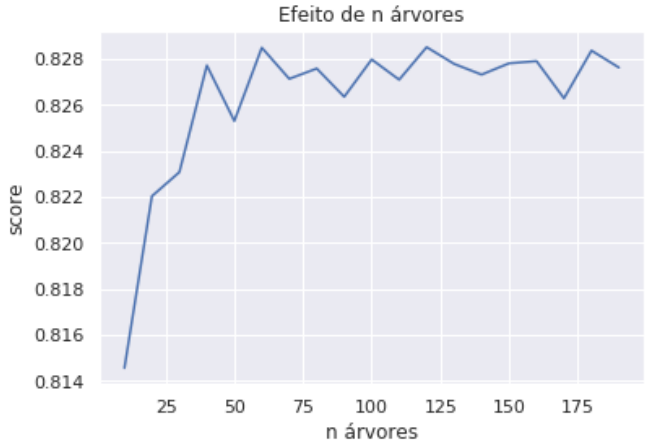


Fig. 4. Gráfico do *score* pelo número de árvores do **Modelo 2**.

B. Regressão por Mínimos Quadrados Parciais

Mínimos Quadrados Parciais (PLS, do inglês *Partial Least Squares*) é um algoritmo de aprendizado supervisionado. É um método que diminui as variáveis independentes para um número menor de componentes não correlacionados. Em seguida, projeta as matrizes das variáveis preditas e também das variáveis observáveis em um novo espaço, onde busca encontrar um modelo de regressão linear nesses componentes ao invés dos dados originais [30].

O PLS permite otimizar o ajuste da curva de regressão através do método dos quadrados mínimos, que busca obter o melhor ajuste para um conjunto de dados tentando minimizar a soma dos quadrados das diferenças entre o valor estimado e os dados observados (tais diferenças são chamadas resíduos) [31]. A regressão com PLS tem a propriedade desejável de que a precisão dos parâmetros do modelo melhora com o número crescente de variáveis e observações relevantes [32]. É frequentemente recomendado quando o foco da pesquisa é a previsão, e não o teste de hipóteses, quando o tamanho da amostra não é grande ou na presença de dados ruidosos [9].

Para este algoritmo não se fez uma etapa explícita separada de calibração e validação. A implementação da função do modelo de PLS (da biblioteca Scikit-Learn) faz iterações até que seja alcançada um valor menor que um valor de tolerância ou até um limite de iterações. Os valores padrões são de 500 iterações com tolerância de $1e-6$.

VI. Experimentos, Resultados e Discussão

COMO visto na Seção III, uma proposta solicitada pelo cliente será analisada pelo **Modelo 1** e, caso obtenha predição acima do limiar estabelecido de aprovação, o processo de classificação é concluído e a proposta é *Aprovada*; caso o *score* predito tenha um valor abaixo do limiar de reprovação, então, também será encerrado o processo de classificação e a proposta será considerada *Reprovada*. Se nenhum dos dois casos anteriores ocorrer, então, a proposta será submetida ao **Modelo 2**, agora com a primeira consulta à *classificação* do cliente realizada junto às entidades externas. Mais uma vez, os casos de aprovação e reprovação são avaliados; não sendo possível classificar, será destinada a uma análise manual para um operador humano classificar.

A base de dados do experimento conta com 96.041 propostas que não foram utilizadas para calibração, treinamento ou validação dos modelos.

A faixa do limiar de aprovação foi 0,5 (inclusive) à 1,0 (exclusive), com incremento de 0,01. Os valores apresentados nas tabelas são um resumo dos resultados, com incremento de 0,05, enquanto os gráficos apresentam os resultados com incremento de 0,01.

A. Experimentos com o Modelo 1

Duas métricas foram usadas para estudo dos limiares: pela acurácia e pelo percentual de propostas classificadas.

1) Avaliação pela Acurácia: A calibração dos valores de limiares foi feita usando a quantidade de acertos, ou a acurácia, de cada modelo usando a base de validação.

A Tabela II exibe as acurácias obtidas para o **Modelo 1** com o algoritmo Floresta Aleatória. A leitura da tabela é feita da seguinte forma: por exemplo, se o limiar de aprovação (coluna "Limiar de Aprovação") de 0,85 for estabelecido (consequentemente, limiar de reprovação de 0,15 na coluna "Limiar de Reprovação"), o acerto de propostas aprovadas e reprovadas foi de 98,05% (coluna "Acurácia M1") no Modelo 1. Na tabela, verifica-se que quanto maior o limiar, maior a acurácia do modelo.

A Tabela III exibe as acurácias obtidas para o **Modelo 1** com algoritmo PLS para a mesma variação de limiares usados para a Floresta Aleatória. A leitura da tabela segue o mesmo princípio da tabela da Floresta Aleatória, por exemplo, se o limiar de aprovação de 0,85 for estabelecido, 98,55% das propostas serão corretamente classificadas no **Modelo 1**. Da mesma forma que no resultado da Floresta Aleatória, verifica-se que quanto maior o limiar no PLS, maior a acurácia. Entretanto, no PLS os valores são menores para o limiar de 0,6 e cresce mais rapidamente que para a Floresta Aleatória, de modo que, que para o último valor de limiar (0,99), o resultado do PLS é um pouco melhor que da Floresta Aleatória.

2) Avaliação pelo Percentual de Propostas: Neste método, foi utilizado como métrica o percentual de propostas que o modelo classifica de acordo com o valor de limiar. Suponha que o limiar de aprovação seja estabelecido em 0,85. Pergunta-se, "Quantas propostas serão classificadas?", e a

TABELA II
Floresta Aleatória - Acurácias obtidas na validação (base **Validação**) do Modelo 1 para diferentes limiares.

Limiar de Aprovação	Limiar de Reprovação	Acurácia M1
0,60	0,40	94,73%
0,65	0,35	95,54%
0,70	0,30	96,22%
0,75	0,25	96,87%
0,80	0,20	97,44%
0,85	0,15	98,05%
0,90	0,10	98,59%
0,95	0,05	99,00%
0,99	0,01	99,33%

TABELA III
PLS - Acurácias obtidas na validação (base **Validação**) do Modelo 1 para diferentes limiares.

Limiar de Aprovação	Limiar de Reprovação	Acurácia M1
0,60	0,40	90,19%
0,65	0,35	92,92%
0,70	0,30	95,22%
0,75	0,25	96,85%
0,80	0,20	97,89%
0,85	0,15	98,55%
0,90	0,10	99,00%
0,95	0,05	99,37%
0,99	0,01	99,53%

TABELA IV
Floresta Aleatória - Percentual do total de classificações obtido nos testes (base **Testes**) do Modelo 1 para diferentes limiares.

Limiar de Aprovação	Limiar de Reprovação	% de Propostas
0,60	0,40	94,83%
0,65	0,35	92,25%
0,70	0,30	89,57%
0,75	0,25	86,94%
0,80	0,20	84,05%
0,85	0,15	80,95%
0,90	0,10	76,95%
0,95	0,05	71,10%
0,99	0,01	58,12%

TABELA V
PLS - Percentual do total de classificações obtido nos testes (base **Testes**) do Modelo 1 para diferentes limiares.

Limiar de Aprovação	Limiar de Reprovação	% de Propostas
0,60	0,40	79,97%
0,65	0,35	71,36%
0,70	0,30	63,23%
0,75	0,25	55,97%
0,80	0,20	48,87%
0,85	0,15	42,14%
0,90	0,10	34,75%
0,95	0,05	26,40%
0,99	0,01	19,31%

resposta mostra quantos registros o modelo classifica sem intervenção manual. Lembrando que a classificação é dada para propostas aprovadas e reprovadas onde o limiar de aprovação limita as propostas aprovadas no limite superior e o limiar de reprovação limita as propostas reprovadas no limite inferior.

Esta metodologia de análise complementa a anterior. Enquanto uma preocupa-se em avaliar dentro de um limiar de aceitação a acurácia do resultado, ou seja, o quanto o modelo está acertando, esta segunda visa mostrar o volume de propostas passíveis de classificação sem intervenção manual dado um limiar. As Tabelas IV e V exibem os resultados para alguns limiares para a Floresta Aleatória e PLS, respectivamente.

A leitura da tabela é feita da seguinte forma: por exemplo, se o limiar de aprovação (coluna "Limiar de Aprovação") de 0,85 for estabelecido (consequentemente, limiar de reprovação de 0,15 na coluna "Limiar de Reprovação"), o percentual de propostas classificadas em aprovadas e reprovadas é de 80,95% (coluna "Percentual de Propostas") no Modelo 1. Na tabela, verifica-se que quanto maior o limiar, menor a quantidade de propostas classificadas. Quanto menor esse percentual, mais propostas estão sendo enviadas para a análise do **Modelo 2**, incorrendo em custos para a financeira.

3) Avaliação pela Acurácia e pelo Percentual de Propostas: Os gráficos das Figuras 5 e 6 mostram a acurácia obtida na classificação versus o volume de propostas classificadas pelos limiares de aprovação e reprovação estabelecidos em cada modelo.

Nestes gráficos, para cada valor de limiar de aprovação exibido no eixo x existe o equivalente (1 - limiar) para as propostas reprovadas. Enquanto os limiares de aprovação são [0,5; 0,6; 0,7; 0,8; 0.9] os de reprovação são [0,5; 0,4; 0,3; 0,2; 0,1]. A curva azul mostra o percentual de acurácia crescente conforme o limiar de aprovação aumenta. A curva vermelha mostra o percentual total de propostas classificadas reduzindo à medida que o limiar de aprovação aumenta.

No gráfico da Floresta Aleatória (Figura 5), para o limiar de aprovação em 0,6 temos que aproximadamente 95% das propostas foram classificadas (aprovadas as que possuem limiar maior que 0,6 e reprovadas as que possuem limiar menor que 0,4) com acurácia de 93%. Quanto maior o limiar de aprovação maior a acurácia e menor é a quantidade de propostas classificadas nesta fase. No gráfico do PLS (Figura 6), para o limiar de aprovação em 0,6 temos que aproximadamente 80% das propostas foram classificadas com acurácia de 90%.

*B. Experimentos com o **Modelo 2***

Os resultados desta segunda fase de análise estão nas Tabelas VI e VII. A leitura desta tabela é diferente das anteriores, pois parte do valor de acurácia final, com os resultados do **Modelo 1** e do **Modelo 2**, e apresenta quais seriam os limiares de cada etapa para se alcançar a acurácia final pretendida. As colunas "Limiar M1" e

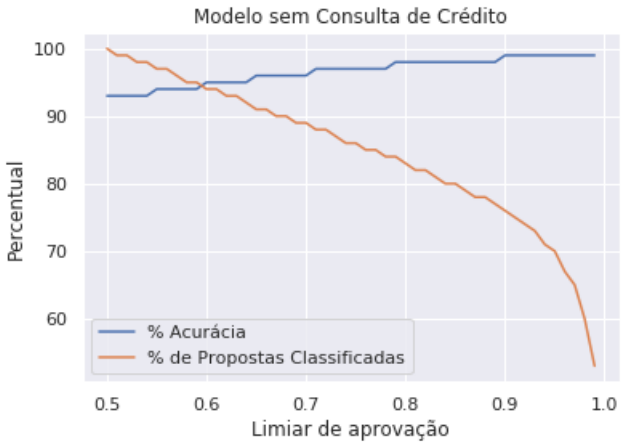


Fig. 5. Floresta Aleatória - Acurácia versus Percentual de Propostas do Modelo 1.

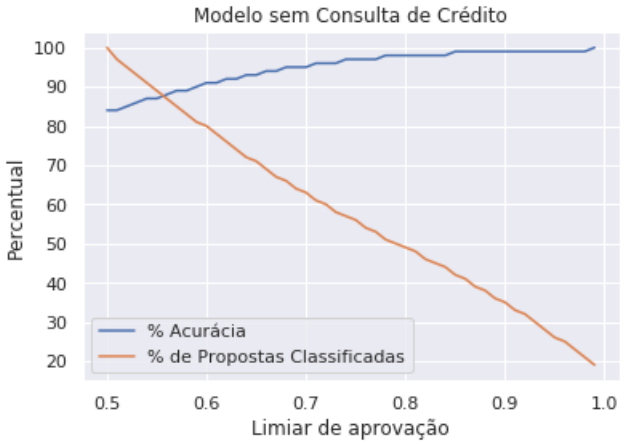


Fig. 6. PLS - Acurácia versus Percentual de Propostas do Modelo 1.

"Limiar M2" referem-se aos limiares necessários para que a acurácia seja satisfeita na classificação dos Modelo 1 e Modelo 2, respectivamente. As colunas "% M1" e "% M2" referem-se aos percentuais do total de propostas classificadas conforme o limiar informado em cada modelo.

Note que quanto maior a acurácia exigida, maior o limiar de aprovação. Para o **Modelo 1** (sem consulta ao crédito) um limiar de 0,75 já é suficiente para prover 97% de acurácia em ambos algoritmos. Porém, pra que essa mesma acurácia se mantenha no **RF Modelo 2** (com uma consulta ao crédito) precisa de 0,99 de limiar de aprovação, enquanto que o **PLS Modelo 2** é necessário um limiar de 0,94.

O algoritmo Floresta Aleatória gerou modelos mais estáveis e consistentes com o que era esperado, principalmente para o Modelo 2. Conforme esperado, ao se adicionar o atributo de *classificação* do cliente houve melhora nos resultados da classificação. Conforme Tabela VI, se for tomada, por exemplo, a acurácia de 97% o limiar de 0,75 do **Modelo 1** classifica 86,56% das propostas e o **Modelo 2** mais 4,04% do restante das propostas não classificadas.

Para o algoritmo PLS observa-se que o **Modelo 2** não se comportou como esperado ao adicionar o atributo de

TABELA VI
Floresta Aleatória - Valores dos limiares necessários em cada modelo para satisfazer a acurácia exigida com os percentuais totais de classificação resultantes.

Acurácia	Limiar M1	Limiar M2	**% M1**	**% M2**
99%	0,96	insuficiente	69,19%	-
98%	0,84	0,999	81,23%	3,64%
97%	0,75	0,99	86,56%	4,04%
96%	0,69	0,98	89,91%	4,4%
95%	0,62	0,98	93,84%	4,42%
94%	0,57	0,902	96,58%	6,65%
93%	0,52	0,82	99,26%	11,24%

TABELA VII
PLS - Valores dos limiares necessários em cada modelo para satisfazer a acurácia exigida com os percentuais totais de classificação resultantes.

Acurácia	Limiar M1	Limiar M2	**% M1**	**% M2**
99%	0,90	0.95	35.33%	14.65%
98%	0,81	0,94	48.06%	5.05%
97%	0,75	0,94	56.12%	4.01%
96%	0,72	0,94	60.41%	2.17%
95%	0,69	0,94	64.08%	0.26%
94%	0,66	0,94	69.58%	0.13%
93%	0,64	impreciso	72.98%	0.04%

classificação do cliente. Houve melhora na classificação final apenas para limiares abaixo de 0,80, acima deste valor a classificação piora mostrando que a nova informação não contribuiu bem com limiares altos. Conforme Tabela VII, por exemplo, para a acurácia de 97% e o limiar de 0,75 do **Modelo 1**, um total de 56,12% das propostas é classificado pelo **Modelo 1** e o **Modelo 2** mais 4,01% do restante das propostas não classificadas. Esse modelo apresentou uma capacidade de classificação inferior ao de Floresta Aleatória e ainda com certa instabilidade em alguns limiares como pode ser observado comparando a coluna **Modelo 1** de ambas as tabelas. Ainda assim, é um resultado melhor que o modelo atual da empresa classificando até 64,08% das propostas se a acurácia for de 95%. A empresa ficou satisfeita com os resultados demonstrados e autorizou as adequações dos procedimentos operacionais para implantação do uso do modelo em ambiente de produção.

A arquitetura provou-se robusta para classificação das propostas com o **Modelo 1** e o **Modelo 2** com resultados mais estáveis com algoritmo Floresta Aleatória.

VII. Conclusões

Este trabalho apresenta uma abordagem baseada em aprendizado de máquina para o processo de classificação de propostas de empréstimos pessoais de uma entidade financeira, bem como de comparar a performance e a adequação destas técnicas à arquitetura proposta. A arquitetura em questão realiza duas etapas de classificação, com dois modelos diferentes, mantendo uma parte da análise de forma manual com o objetivo de preservar os empregos dos colaboradores da financeira. O resultado foi um sistema que apresentou acurácia melhor que o sistema existente, e que diminuísse o custo de acesso aos dados de órgãos de proteção ao crédito.

Iniciamos a análise exploratória dos dados carregando um dataset de quase 200.000 propostas desde janeiro de 2018 até abril de 2019, contendo informações do perfil do cliente, histórico de relacionamento do cliente com a financeira e dados da proposta. A partir daí, removeram-se os *outliers*, os casos incoerentes, removeram-se atributos com menor representatividade estatística, removeu-se atributos que possuíam alto coeficiente de correlação de Pearson e, por fim, dividiu-se o dataset tratado em duas grandes partes de aproximadamente 96 mil propostas cada, para treinamento e testes, respectivamente.

Foram escolhidas duas técnicas de aprendizado de máquina, os algoritmos de Floresta Aleatória e o de PLS. Para construir os modelos, o primeiro esforço se concentrou em identificar um número ideal de árvores da Floresta Aleatória. Após a calibração do modelo, foi definido 100 árvores para o Modelo 1 e 60 árvores para o Modelo 2. Para o PLS utilizamos os parâmetros padrão da biblioteca usada.

Os resultados contínuos dos regressores foram convertidos em resultados discretos, avaliando limiares de aprovação e reprovação. Os experimentos foram analisados em relação a duas métricas: a primeira mediu a acurácia dos resultados sob de um limiar de certeza esperado e a segunda mediu o volume de propostas que o modelo foi capaz classificar sob o limiar de aprovação esperado. Foi visto que, quanto maior o limiar, maior a acurácia e menor o volume de propostas classificadas.

O processo proposto nos experimentos apresentou resultados muito interessantes. Além de classificar com alta acurácia também classificou considerável volume de propostas. O atual modelo utilizado na empresa classifica automaticamente aproximadamente 52% das propostas. Para a acurácia mais alta de 99% (que implica na menor quantidade de propostas avaliadas) a arquitetura proposta, com o Floresta Aleatória, já classificaria 69,19%, um ganho substancial. Assim, há uma redução considerável de propostas para a próxima etapa, e com isso, há uma redução direta com custos de consulta aos órgãos de crédito, uma vez que a maior parte das propostas é classificada pelo modelo sem consulta. Apenas 30,81% (para acurácia 99%) das propostas consultariam o crédito do cliente enquanto o modelo atual consulta o crédito em cerca de 48% das propostas.

Constata-se que a consulta ao crédito realizada hoje em muitos casos é desnecessária, uma vez que o modelo sem consulta conseguiu reproduzir o resultado existente, o que significa que a informação junto ao órgão de proteção ao crédito não tem tanto peso na informação sobre classificação de crédito do cliente quanto a financeira utiliza. Manteve-se uma parte das propostas sob análise manual por operadores de crédito, que é importante para a manutenção dos empregos existentes. Com o estudo dos limiares, caberá à empresa responder qual o limiar aceitável para o modelo operar.

É importante ressaltar que não se considerou neste trabalho avaliar a qualidade do crédito concedido, mas apenas desenvolver modelos que representem bem o atual comportamento da empresa ao avaliar uma proposta de crédito. Isso significa que o sistema proposto buscou literalmente aprender como a empresa classifica as propostas de crédito e não entrou no mérito se essa classificação é boa ou ruim. Modelos que avaliem a qualidade do crédito e sua consequente taxa de inadimplência ficam para um possível trabalho futuro.

É importante que a financeira faça o acompanhamento do sistema para avaliar, no futuro, quantos empréstimos concedidos se tornaram inadimplentes. E também avaliar a mudança em sua base de dados e no sistema para armazenar os dados históricos da renda de seus clientes.

Trabalhos futuros poderiam avaliar outras técnicas de regressão, calibração dos hiper-parâmetros dos modelos, uso de metodologias de validação cruzada avaliação e análise do sistema com propostas reais no ambiente de produção. Considerando que existem duas avaliações baseadas no limiar que são inversas (o aumento de uma causa a diminuição de outra), é interessante verificar outras formas de cálculo para os limiares. Outras métricas de avaliação, tais como precisão, revocação e medida-F1, também devem ser calculadas e analisadas. Também planeja-se realizar uma análise da existência ou não de variáveis endógenas dentre as utilizadas para construção dos modelos e, se for o caso, avaliar o impacto da endogeneidade nos resultados dos modelos construídos.

Referências

[1] R. S. da Paixão, A. A. Herzog e S. M. Casagrande, "Um Estudo das Instituições Financeiras e as suas Principais Linhas se Créditos de Curto Prazo Ofertadas do Brasil", *Revista Científica Fetes*, v. 1, n. 1, pp. 3–18, 2019.

[2] M. Gertler e P. Karadi, "Monetary policy surprises, credit costs, and economic activity", *American Economic Journal: Macroeconomics*, v. 7, n. 1, pp. 44–76, 2015.

[3] UOL, *BC autoriza 1ª fintech a conceder empréstimo no país sem mediação de banco*, https://economia.uol.com.br/noticias/redacao/2018/12/05/banco-central-concede-autorizacao-para-primeira-fintech\-de-credito.htm?cmpid=copiaecola, Accessed on 01/05/2019, dez. de 2018.

[4] P. Zogbi e M. Castro, *Empresa Simples de Crédito (ESC): entenda a lei que permite empréstimos entre pessoas*, https://www.infomoney.com.br/minhas-financas/credito/noticia/8088402/esc-como-vai-funciona-a-empresa-criada-para-facilitar\-a-vida-de-pmes, Accessed on 01/08/2019, jun. de 2019.

[5] F. de Oliveira Araújo e A. P. A. Ribeiro, *Mercado de crédito brasileiro*, 1ª. Organização Internacional do Trabalho, 2003.

[6] S. Kealhofer, "Quantifying credit risk I: default prediction", *Financial Analysts Journal*, v. 59, n. 1, pp. 30–44, 2003.

[7] F. Louzada, A. Ara e G. B. Fernandes, "Classification methods applied to credit scoring: Systematic review and overall comparison", *Surveys in Operations Research and Management Science*, v. 21, n. 2, pp. 117–134, 2016, ISSN: 18767354. DOI: 10.1016/j.sorms.2016.10.001. arXiv: 1602.02137. endereço: http://dx.doi.org/10.1016/j.sorms.2016.10.001.

[8] L. Breiman, "Random forests", *Machine learning*, v. 45, n. 1, pp. 5–32, 2001.

[9] G. D. Garson, "Partial least squares: Regression and structural equation models", *Asheboro, NC: Statistical Associates Publishers*, 2016.

[10] M. S. d. Vasconcellos, "Proposta de método para análise de concessões de crédito a pessoas físicas", tese de dout., Universidade de São Paulo, 2002.

[11] P. Renton, *Peer To Peer Lending Crosses $1 Billion In Loans Issued*, https://techcrunch.com/2012/05/29/peer-to-peer-lending-crosses-1-billion-in-loans-issued/, Acessado em 2020-04-20, Techcrunch, mai. de 2012. (acesso em 20/04/2020).

[12] D. S. Rodrigues, A. R. A. Brasil, M. B. Costa, K. S. Komati e L. A. Pinto, "Uma Análise Comparativa de um Algoritmo de Aprendizado Supervisionado para Solicitações de Empréstimo em uma Plataforma Peer-to-Peer", em *Anais do XIV Simpósio Brasileiro de Sistemas de Informação*, Caxias do Sul: SBC, 2018, pp. 332–325. endereço: https://sol.sbc.org.br/index.php/sbsi/article/view/5104.

[13] M. Polena e T. Regner, "Determinants of borrowers' default in P2P lending under consideration of the loan risk class Jena", Friedrich Schiller University Jena, Jena, Jena Economic Research Papers, No. 2016-023, 2016, p. 30.

[14] Y. Zhang, H. Li, M. Hai, J. Li e A. Li, "Determinants of loan funded successful in online P2P Lending", *Procedia Computer Science*, v. 122, pp. 896–901, 2017, ISSN: 18770509. DOI: 10.1016/j.procs.2017.11.452. endereço: https://doi.org/10.1016/j.procs.2017.11.452.

[15] G. Wang, J. Ma, L. Huang e K. Xu, "Two credit scoring models based on dual strategy ensemble trees", *Knowledge-Based Systems*, v. 26, pp. 61–68, 2012, ISSN: 09507051. DOI: 10.1016/j.knosys.2011.06.020. endereço: http://dx.doi.org/10.1016/j.knosys.2011.06.020.

[16] I. Brown e C. Mues, "An experimental comparison of classification algorithms for imbalanced credit scoring data sets", *Expert Systems with Applications*, v. 39, n. 3, pp. 3446–3453, 2012, ISSN: 09574174. DOI: 10.1016/j.eswa.2011.09.033. endereço: http://dx.doi.org/10.1016/j.eswa.2011.09.033.

[17] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms", *Neural computation*, v. 10, n. 7, pp. 1895–1923, 1998.

[18] A. C. da Silva e B. C. X. Bianca, "Inadimplência: Um estudo com usuários de cartão de crédito em Belo Horizonte/MG", *e3*, v. 4, n. 2, pp. 86–110, 2018.

[19] J. E. Chen e C. W. Hsiang, "Causal random forests model using instrumental variable quantile regression", *Econometrics*, v. 7, n. 4, pp. 1–22, 2019, ISSN: 22251146. DOI: 10.3390/econometrics7040049.

[20] M. Biggs, "Prescriptive analytics in operations problems: a tree ensemble approach", PhD Thesis, Massachusetts Institute of Technology, 2019.

[21] G. E. d. A. P. A. Batista, "Pré-processamento de dados em aprendizado de máquina supervisionado, Orientadora: Maria Carolina Monard", Doutorado em Ciências, Universidade de São Paulo, São Carlos, 2003, p. 204.

[22] I. Guyon e A. Elisseeff, "An introduction to feature extraction", em *Feature extraction*, Springer, 2006, pp. 1–25.

[23] R. Simon, "Resampling strategies for model assessment and selection", em *Fundamentals of data mining in genomics and proteomics*, Springer, 2007, pp. 173–186.

[24] J. M. Zhang, M. Harman, L. Ma e Y. Liu, "Machine learning testing: Survey, landscapes and horizons", *arXiv preprint arXiv:1906.10742*, 2019.

[25] S. Ali, S. Adnan, T. Nawaz, M. O. Ullah e S. Aziz, "Human heart sounds classification using ensemble methods", *University of Engineering and Technology Taxila. Technical Journal*, v. 22, n. 1, p. 113, 2017.

[26] F. G. Furat e T. Ibrikci, "Classification of Down Syndrome of Mice Protein Dataset on MongoDB Database", *Balkan Journal of Electrical and Computer Engineering*, v. 6, n. 2, pp. 112–117, 2018.

[27] X. Zhuang, A. Ni, L. Liao, Y. Guo, W. Dai, Y. Jiang, H. Zhou, X. Hu, Z. Du, X. Wang et al., "Environment-wide association study to identify novel factors associated with peripheral arterial disease: Evidence from the National Health and Nutrition Examination Survey (1999–2004)", *Atherosclerosis*, v. 269, pp. 172–177, 2018.

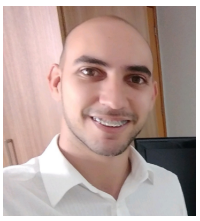
[28] H. Aguilera, C. Guardiola-Albert e C. Serrano-Hidalgo, "Estimating extremely large amounts of missing precipitation data", *Journal of Hydroinformatics*, v. 22, n. 3, pp. 578–592, 2020.

[29] T. K. Ho, "Random decision forests", em *Proceedings of 3rd international conference on document analysis and recognition*, IEEE, vol. 1, 1995, pp. 278–282.

[30] T. D. Science, *Key Types of Regressions: Which One to Use?*, https://towardsdatascience.com/key-types-of-regressions-which-one-to-use-c1f25407a8a4, Accessed on 01/10/2020, mai. de 2019.

[31] P. Geladi e B. R. Kowalski, "Partial least-squares regression: a tutorial", *Analytica chimica acta*, v. 185, pp. 1–17, 1986.

[32] S. Wold, M. Sjöström e L. Eriksson, "PLS-regression: a basic tool of chemometrics", *Chemometrics and intelligent laboratory systems*, v. 58, n. 2, pp. 109–130, 2001.



Pablo Simões Nascimento possui graduação em Ciência da Computação pela Universidade Federal do Espírito Santo (2012) e pós-graduação em Ciência de Dados com Big Data pelo IFES Serra (2019). Trabalha com desenvolvimento de software desde 2009 para empresas no setor privado. Atuou como desenvolvedor e analista de software para sistemas de informação e serviços da área financeira e de análise de crédito por 7 anos. Trabalhou com tecnologias .Net em sistemas Web, desktop e gerenciamento e migração de bancos de dados. Atualmente trabalha como professor de Algoritmos e Lógica de programação na Faculdade Pitágoras e como analista de software na empresa Biancogres, destaque nacional na fabricação de porcelanatos e revestimentos.

Karin S. Komati É professora do Instituto Federal do Espírito Santo desde 2012 e atual Coordenadora do Mestrado em Computação Aplicada do Campus Serra. Possui formação acadêmica com graduações em: bacharelado em Ciência da Computação pela UFES (1995), e graduação em Engenharia Elétrica pela UFES (1997). Estas duas áreas se refletem na pós-graduação, pois é Doutora em Engenharia Elétrica pela UFES (2011) e é Mestre em Informática pela UFES (2002). Atua em docência do ensino superior desde 1998, trabalhando em diversas instituições privadas e públicas. Já atuou como analista de sistemas da empresa multinacional Xerox (1994-1998) e sócia-proprietária de micro-empresa de prestação de serviços em desenvolvimento de sistemas (1999-2003). No ano de 2006, trabalhou em desenvolvimento Web, na empresa Softcreate no Japão. A área de pesquisa se concentra em Processamento Digital de Imagens, Reconhecimento de Padrões e Banco de Dados. É líder do grupo Nu[Tec] (http://dgp.cnpq.br/dgp/espelhogrupo/36297). Foi Diretora de Pesquisa, Pós-graduação e Extensão por mais de 3 anos, responsável pelo Núcleo Incubador do Campus Serra (NIS), depois foi coordenadora de pesquisa e liderou as duas propostas de novos cursos de pós-graduação "Mestrado Profissional em Engenharia de Controle e Automação"submetida à CAPES em 2014 e aprovada na 155ª reunião do CTC-ES da CAPES e o "Mestrado Profissional em Computação Aplicada"submetida à CAPES em 2017 e aprovada na 179ª reunião do CTC-ES da CAPES.

Jefferson Oliveira Andrade Recebeu o título de Engenheiro de Computação em 1995, e o título de Mestre em Informática em 2001, ambos pela Universidade Federal do Espírito Santo. Ele possui vários anos de experiência como líder de equipes em projetos de desenvolvimento de software, tanto em empresas locais quanto multinacionais no Brasil. De 2005 a 2008 foi membro do Programming Logic Group, na Universidade de Tsukuba, no Japão. Em 2013 recebeu seu Doutorado em Educação pela Universidad del Norte, no Paraguai (revalidado pela UFPR em 2016), pela sua pesquisa sobre a aplicação de gamificação no ensino de lógica formal a alunos de graduação do curso de Sistemas de Informação. Atualmente o Dr. Andrade é professor titular do Campus Serra do Instituto Federal do Espírito Santo, onde faz parte do Programa de Pós-graduação em Computação Aplicada. Seus interesses de pesquisa incluem aplicações de ciência de dados, métodos formais de desenvolvimento de software, verificação formal de sistemas, verificação de modelos, lógicas multi-valoradas e probabilísticas, ensino de lógica e de métodos formais.