Modelo Preditivo para Avaliação de Crédito em Empréstimos Pessoais

Pablo Simões Nascimento[1], **Karin S. Komati**[1] **Jefferson O. Andrade**[1],

[1] Campus Serra – Instituto Federal do Espírito Santo
Rodovia ES-010 – Km 6,5 – Manguinhos – Serra – ES – Brazil

pablosnascimento@gmail.com, {jefferson.andrade,kkomati}@ifes.edu.br

Abstract. *This work consists in the construction of a classification model for credit proposal analysis using machine learning algorithms from the historical database of credit analysis made by a financial institution. The study uses a database of nearly 200.000 proposals, performs exploratory data analysis, generates two prediction models, sets thresholds for pass and fail proposals, tests initial scores to obtain 0.77 and 0.83 for models 1 and 2 respectively, It establishes the accuracy and quantity of proposals classified as model evaluation methodologies, performs experiments varying the thresholds and presents the results. Assuring accuracy of 97% we were able to correctly classify 90.6% of proposals.*

Resumo. *Este trabalho consiste na construção de um modelo de classificação para análise de propostas de crédito utilizando algoritmos de aprendizado de máquina a partir da base de dados histórica de análises de crédito feitas por uma instituição financeira. O estudo utiliza uma base de dados com quase 200.000 propostas, realiza a análise exploratória dos dados, gera dois modelos de predição, estabelece limiares para aprovação e reprovação das propostas, testa os scores iniciais obtendo 0,77 e 0,83 para os modelos 1 e 2 respectivamente, estabelece a acurácia e quantidade de propostas classificadas como metodologias de avaliação do modelo, realiza experimentos variando os limiares e apresenta os resultados. Garantindo acurácia de 97% pudemos classificar corretamente 90,6% das propostas.*

1. Introdução

A concessão de crédito consiste em um mecanismo para o aquecimento da economia, uma vez que aumenta o poder de consumo das pessoas, fomenta a produção industrial e incentiva expansão de oferta de serviços [Gertler and Karadi 2015]. Tais recursos possibilitam, por exemplo, que pessoas realizem seus projetos pessoais, melhorem sua qualidade de vida através da aquisição de bens de consumo e, no caso das empresas, que estas invistam na expansão e melhoria de seu parque produtivo, de seus processos ou recursos humanos.

Até há pouco tempo no Brasil, apenas empresas de sociedade de crédito autorizadas pelo Banco Central poderiam fazer empréstimos, basicamente os bancos. 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 estão autorizadas para prestar concessões de créditos sem a mediação de um banco [UOL 2018].

Mesmo uma pessoa física pode abrir uma ESC (Empresa Simples de Crédito) e conceder empréstimos [Zogbi and Castro 2019].

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. Para uma pessoa física obter um empréstimo há duas formas mais comuns [de Oliveira Araújo and Ribeiro 2003]: 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, muitos mutuários deixam de realizar o pagamento das parcelas e se tornam inadimplentes [Kealhofer 2003]. 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 prestadas pela mesma, e suas condições de honrar os compromissos financeiros e não entrar 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. Somente em 2012 por meio de novo planejamento estratégico passou a fomentar forte posição no mercado chegando 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 já configuradas previamente (conforme a experiência de um gerente de crédito e de alguns dados estatísticos históricos) e pode dar 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; esta é a segunda etapa chamada de análise manual ou mesa de crédito. Atualmente, a primeira fase, a de análise automática, responde a 52% das propostas. Os outros 48% são analisados manualmente.

Este trabalho busca construir um modelo para análise automática de propostas de crédito que seja mais assertivo tornando 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.

Para tanto, utilizaremos a base de dados interna de clientes, o histórico de propostas analisadas, informações sobre o cliente em órgãos de crédito, técnicas de pré-processamento e limpeza dos dados, algoritmos de aprendizado de máquina para desenvolvimento do modelo bem como ferramentas de visualização de dados e verificação dos

resultados para conclusões e validação do modelo encontrado.

2. Trabalhos Relacionados

Muitos trabalhos relacionados à procura por uma melhor metodologia para a análise de crédito pessoal têm sido desenvolvidos. Em Vasconcellos [2002], aborda-se uma proposta de metodologia 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 ele trabalha. 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.

O autor difere deste trabalho principalmente com relação ao critério utilizado para classificação da proposta de crédito que foi a inadimplência. Enquanto propomos uma abordagem de classificação alternativa à utilizada atualmente e baseada em um histórico de propostas já classificadas, o autor propôs um modelo tal, também baseado no histórico de propostas recentes, porém, focando na qualidade dos créditos concedidos cujo resultado apresenta um indicador de tendência à inadimplência.

O modelo proposto pelo autor não se preocupou, portanto, em replicar o comportamento da atual análise de crédito feita pela instituição. Ao invés disso, buscou melhorar tal modelo oferecendo um que seja capaz de classificar melhor a proposta pretendendo evitar alto índice de inadimplência. 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 que ainda que a taxa de inadimplência seja reduzida como resultado de uma boa análise, reconhece que isso 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.

Um outro trabalho que buscou prover metodologias capazes de melhorar a análise de crédito foi desenvolvido por [Rodrigues et al. 2018]. 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 como trabalhos comparativos.

A pesquisa difere deste trabalho em dois pontos principais: o primeiro, semelhante ao de Vasconcellos, 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 o trabalho buscou desenvolver análises comparativas dentre vários algoritmos de classificação buscando evidenciar qual melhor se aplica à predição de inadimplência, o que não se aplicou a este trabalho.

3. Materiais e Métodos

Um modelo preditivo 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 [Dietterich 1998]. A natureza do problema discutido neste trabalho é de classificação, uma vez que cada proposta tem resultado binário sendo aprovada ou reprovada. Porém, um modelo de classificação não seria aplicável à realidade da empresa foco desta solução pelo fato de excluir a necessidade de qualquer análise manual ao classificar todas as propostas. Desta maneira todo o setor de crédito e seus analistas perderiam a função. Por isso aplicamos a este trabalho o estimador do tipo regressor que nos permite definir limites acima e abaixo dos quais classificaremos as propostas e as que não forem classificadas continuarão demandando o setor de crédito com análise manual.

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

No entanto, há propostas que podem estar entre estes limiares, e segue para uma segunda fase. Passa-se a consulta aos órgãos de crédito, ou seja, são externos aos dados proprietários da empresa. Nem todos os clientes possuem tais informações, e existe um custo financeiro para cada consulta feita aos órgãos de crédito. A próxima etapa possui um modelo, o qual chamaremos **Modelo 2**. Tal arquitetura permite consultar a situação do cliente apenas quando necessário, ou seja, caso a proposta não tenha uma confiança suficiente para defini-la como aprovada ou reprovada, diminuindo o gasto financeiro na busca de tal informação.

3.1. Base de dados

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 cancelada é quando a pessoa iniciou o processo de solicitação de crédito, mas não avançou na entrada de dados e mesmo após um contato por telefone não quis continuar com o processo, 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. O Pré-aprovado é quando há uma campanha de marketing para que pessoas possam ser novos clientes da financeira, mas que ainda não efetivaram concessão de crédito. Para o propósito deste trabalho apenas propostas Aprovadas ou Reprovadas foram selecionadas.

Uma consideração sobre os dados é quanto à sua temporalidade. Alguns dados cadastrais dos clientes são atualizados apenas quando uma nova proposta de crédito é solicitada, o valor do salário ou renda, por exemplo. 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

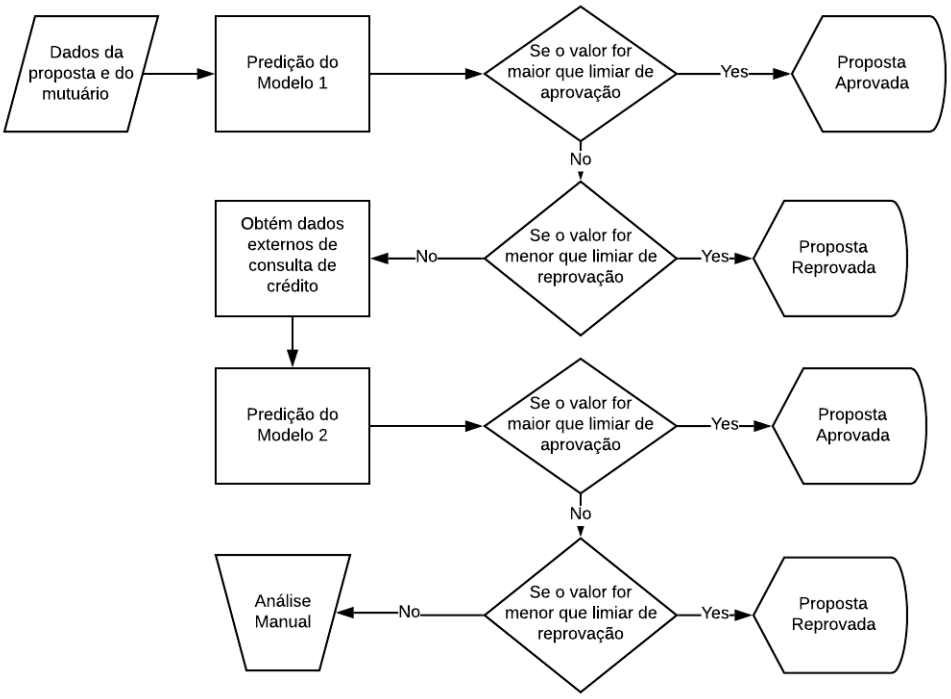

Figura 1. Fluxo do processo proposto para análise das propostas de crédito.

limitação uma condição foi necessária no filtro de todas as propostas: apenas as propostas de clientes que não tiveram cadastro atualizado foram selecionadas. Assim garantimos 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.

Por fim, um total de 48 atributos em quase 200.000 propostas entre janeiro de 2018 e março de 2019 foram selecionadas. Tais atributos estão distribuídos conforme a Tabela 1.

3.1.1. Análise exploratória e Pré-processamento dos dados

De acordo com Batista [2003], 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 feita utilizando a linguagem Python, com pacotes Pandas, *Scikit Learn* e numpy, na plataforma do Jupyter Notebook.

Observamos que nossa base de dados é razoavelmente bem balanceada possuindo 62% das propostas aprovadas e 38% reprovadas. Alguns testes iniciais feitos buscando balancear a base para 50% provocaram certa perda na acurácia da predição final, portanto mantivemos o percentual original.

Alguns dados precisaram ser tratados, como a remoção de *outliers*, tratamento dos valores nulos e enriquecimento de semântica pela criação derivada de novos dados.

Tipo de Informação	Total de Atributos
Dados cadastrais (Idade, CPF divergente da UF de residência, Tipo de atividade, Reside em UF diferente, Classificação do cliente, Situação conjugal, Tempo de serviço, Localidade, UF de residência, Valor do salário, Valor de outras rendas, Valor de salário do cônjuge, Valor total da renda, Sexo, Código do cliente, Tipo de cliente: Novo ou Recompra)	16
Histórico de relacionamento (Fora da praça de atuação, Fiador em atraso, Parcelas atrasadas a mais de 30 dias, Parcelas atrasadas entre 15 e 30 dias, Quantidade de acordos de cobrança em aberto, Quantidade de contratos em fraude, Percentual quitado do primeiro contrato, Possui pendência jurídica, Quantidade contratos não quitados, Quantidade de propostas anteriores aprovadas, Margem bruta, Percentual quitado último contrato, Cliente sem contrato em aberto, Quantidade contratos quitados antes da proposta, Quantidade contratos antes da proposta, Quantidade de propostas analisadas recentemente, Quantidade de acordos)	17
Dados da proposta (Código do estabelecimento, Tipo de operação: CP ou CDC, Condição comercial, 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 biometria, Data e hora da proposta, Situação da proposta)	15

Tabela 1. Descrição dos atributos, organizados por grupo de dados.

A base de dados inicial continha várias inconsistências, e as mesmas foram retiradas, tais como registros com valores nulos e cadastro de clientes menores que 18 anos. Além disso, todos os valores nulos dos atributos foram substituídos zero.

A data da proposta possui informação de dia e hora em que o cliente a solicitou. Após alguns testes utilizando os valores de data quebramos a data em vários campos de informação mais granulares identificando o dia, mês, ano e hora de forma independente. Esse ajuste deu uma melhora de acurácia no modelo de 2% do que usando apenas dia e mês. Interessante foi o fato constatado de que propostas que são geradas pela manhã tem maior chance de serem aprovadas que as feitas à tarde.

Identificou-se todos os atributos de tipo não numérico e como estavam as frequências de seus valores nos registros. O atributo *forma de pagamento* possuía basicamente a mesma forma em todos os registros (Boleto 99,3% e Cheque 0,7%) e foi removido da análise. Os outros atributos não numéricos possuem valores bem distribuídos, então, todos 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.

A remoção de *outliers* foi realizada ao verificar-se alguns valores muito discrepantes. Haviam valores para os atributos "valor a financiar" e "valor da renda" extremamente

atípicos, e os mesmos foram excluídos da base de dados.

3.1.2. Seleção de Atributos

A seleção de atributos foi feita com base no coeficiente de Pearson [Guyon and Elisseeff 2006]. A decisão de projeto foi manter apenas os atributos cuja correlação entre si fosse menor que 95%, do contrário julgamos redundantes a informação. 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.

Outra forma de análise foi pela variedade de ocorrências de cada valor para todos atributos restantes. Assim, identificou-se características sem representatividade, onde quase todos registros possuem o mesmo valor. Segundo esse critério mais 5 características foram excluídas da base de dados.

Ao final, a base de dados ficou com 40 atributos e 192.177 registros e foi dividida igualmente em duas partes: a primeira para treino e validações do modelo e a segunda para os experimentos.

3.1.3. Dados de órgãos de crédito

As informações de crédito externas são fornecidas por órgãos de crédito como SPC, Boa Vista ou Serasa. 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" ou "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 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*.

3.2. Modelo Preditivo

Floresta Aleatória (do inglês *Random Forest*) é um algoritmo de aprendizagem supervisionado [Ho 1995]. É um método *Ensemble* do tipo *Bagging*, isto é, que constrói vários modelos em paralelo (árvores de decisão) a partir de diferentes sub amostras 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, em paralelo. 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 à 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 do *Random Forest* pode ser utilizado tanto para regressão quanto para classificação. Para o caso de regressão uma especialização deste algoritmo é necessária. Chamado *Random Forest Regressor* este método fornece como saída um valor contínuo de predição. Enquanto a classificação constrói modelos preditivos para valores discretos a regressão é utilizada para valores contínuos.

3.2.1. Treinamento e Calibração dos Modelos de Predição

Para a construção de nosso modelo de predição iniciamos filtrando nosso dataset para 30.000 registros, selecionadas aleatoriamente, devido ao tempo custoso de processamento para treinar o modelo muitas vezes. Para todos os modelos usou-se o método *holdout*, dividiu-se a base em duas partes de treino e testes: 70% para treino e os outros 30% para testes.

O algoritmo de treino do modelo *Random Forest Regressor* é sensível a dados não normalizados, logo, precisamos normalizá-los. Para esta tarefa a classe *StandardScaler* da biblioteca *Scikit Learn* foi utilizada. O que este método faz é transformar o dado para média próxima de zero e um desvio padrão próximo a um e assumindo que não temos valores discrepantes nos dados normaliza-os. Para o treinamento 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* [Rodrigues et al. 2018]. O gráfico da Figura 2 mostra o resultado do *score* pela quantidade de árvores. 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 dos 2 atributos de informação extra sobre a *classificação* do cliente. Somente propostas com *classificação* devem ser utilizadas para o treinamento deste modelo. Do contrário, estaríamos treinando o modelo com os mesmos dados do anterior o que geraria um erro no resultado. Assim, para o **Modelo 2**, filtrou-se a base de dados com registros que possuíssem uma consulta. O gráfico da Figura 3 mostra que a partir de 75 árvores o crescimento do score não melhora muito o modelo, assim, o Modelo 2 ficou com 75 árvores.

3.3. Definição dos limiares

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:

Proposta é **aprovada** se predição >= limiar
Proposta é **reprovada** se predição <= (1 - limiar)

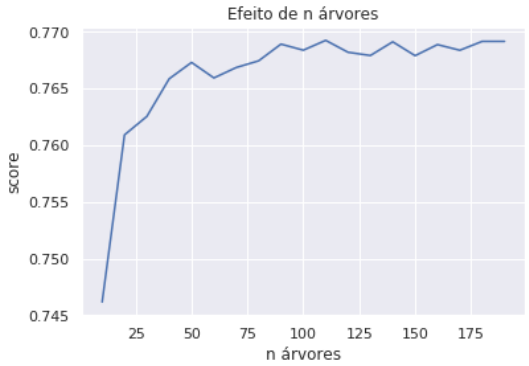

Figura 2. Modelo 1

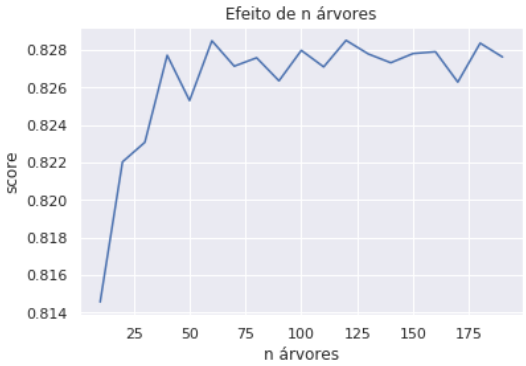

Figura 3. Modelo 2

Portanto, quando falamos em *limiar de aprovação* estamos nos referindo ao limite superior acima do qual as propostas são aprovadas e *limiar de reprovação* sendo o limite inferior equivalente (1 - limiar de aprovação) abaixo do qual as propostas são reprovadas.

Duas abordagens foram tomadas para definição dos limiares: um mede a acurácia e outro a capacidade dos modelos.

3.3.1. Medindo a acurácia do modelo

A calibração deste valor de limiar foi feita usando a metodologia que busca do melhor acerto de acurácia na base de testes. Assim, para um determinado valor de limiar de aprovação e reprovação, computa-se o acerto.

Limiar Ap.	Limiar Rep.	Acurácia M1	Acurácia M2
0,60	0,40	94,73%	95,99%
0,65	0,35	95,54%	96,64%
0,70	0,30	96,22%	97,09%
0,75	0,25	96,87%	97,57%
0,80	0,20	97,44%	98,04%
0,85	0,15	98,02%	98,57%
0,90	0,10	98,59%	98,93%
0,95	0,05	99,00%	99,18%
0,99	0,01	99,33%	99,37%

Tabela 2. Acurácias obtidas nos testes dos Modelo 1 e Modelo 2 para cada limiar.

A tabela acima exibe as acurácias obtidas para os Modelo 1 e 2 com os limiares estabelecidos. Por exemplo, se o limiar de aprovação de 0,85 for estabelecido, mais de 98% das propostas serão corretamente classificadas em ambos os modelos.

3.3.2. Medindo a capacidade do modelo

Neste método, foi utilizado como o valor do limiar de aprovação o percentual de propostas que o modelo é capaz de classificar. Imagine que queremos resultados com limiar

acima de 0,85. A pergunta é "Quantas propostas serão classificadas?", e a resposta mostra o quanto o modelo é capaz de processar. 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 nosso modelo está acertando e próximo dos resultados reais, esta segunda visa mostrar o volume de propostas que o modelo é capaz de classificar. A próxima tabela exibe os resultados para alguns limiares estabelecidos.

Limiar Ap.	Limiar Rep.	Classificações M1	Classificações M2
0,60	0,40	94,83%	97,00%
0,65	0,35	92,25%	95,25%
0,70	0,30	89,57%	93,45%
0,75	0,25	86,94%	91,48%
0,80	0,20	84,05%	89,21%
0,85	0,15	80,95%	86,64%
0,90	0,10	76,95%	82,77%
0,95	0,05	71,1%	77,25%
0,99	0,01	58,12%	66,49%

Tabela 3. Percentual do total de classificações obtido nos testes dos Modelo 1 e Modelo 2 para cada limiar.

A seguir tem-se as matrizes de confusão com os valores de saída verdadeiro e o predito, na Figura 4. Tal visualização nos permite observar a precisão da predição dos modelos obtida nos testes para o limiar de 0,85.

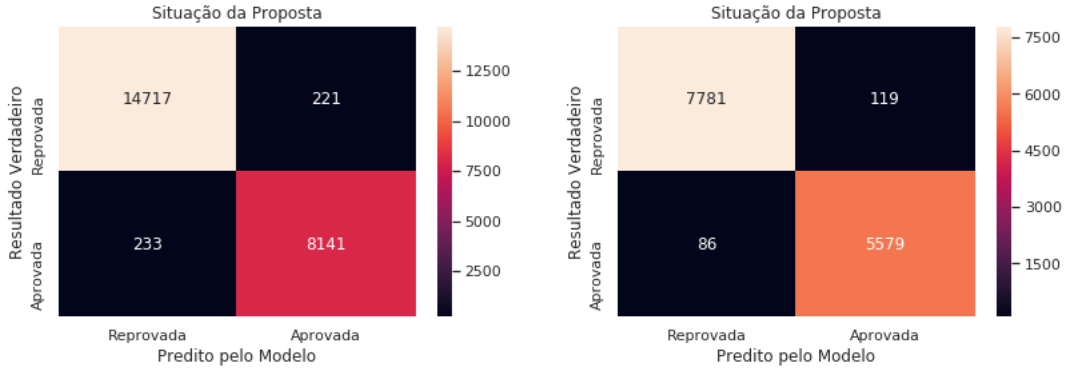

Figura 4. Matriz de confusão para cada modelo com limiar de 0,85

Quanto ao volume de propostas classificadas, se o limiar de 0,85 for estabelecido, para o **Modelo 1** 80,95% das propostas foram classificadas. No **Modelo 2**, 86,64% foram classificadas.

4. Experimentos, Resultados e Discussão

Os experimentos seguiram o fluxo proposto: a proposta solicitada pelo cliente é avaliada pelo motor de crédito (sistema) onde resulta em aprovada, reprovada ou repassada para análise manual.

Relembrando que a proposta solicitada pelo cliente será analisada pelo Modelo 1, caso obtenha predição acima do limiar estabelecido de aprovação, fim do processo e a proposta é Aprovada; caso tenha valor abaixo do limiar de reprovação, então, é considerada proposta Reprovada. Se não entrar em nenhum dos casos anteriores, então será submetida ao Modelo 2, agora com a primeira consulta à classificação do cliente. Mais uma vez, os casos de aprovação e reprovação são avaliados; do contrário será destinada a uma análise manual para um operador de crédito classificar.

4.1. Experimentos com o Modelo 1

A base de dados do experimento conta com 96.041 propostas que não foram utilizadas para treinamento nem testes dos modelos.

O gráfico da Figura 5 mostra a acurácia obtida na classificação e o volume de propostas classificadas pelo limiar de aprovação estabelecido. A curva azul mostra um 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, 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].

Por exemplo, 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%. Naturalmente, quanto maior o limiar de aprovação maior a acurácia e menos propostas são classificadas nesta fase.

4.2. Experimentos com o Modelo 2

Para esta etapa, que é a segunda fase de análise, os resultados estão na Tabela 4. A tabela mostra qual o limiar necessário de cada modelo para que a acurácia seja a mesma ao final da classificação. A tabela exibe os resultados após as propostas terem sido submetidas aos modelos conforme fluxo exposto anteriormente. As colunas Limiar M1 e Limiar M2 referem-se aos limiares necessários para que a acurácia seja satisfeita na classificação dos Modelo 1 e Modelo 2 respectivamente. As colunas Modelo 1 e Modelo 2 referem-se aos percentuais do total de propostas classificadas conforme a acurácia e o limiar informados 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 0,84 já é suficiente para prover 98% de acurácia. Porém, pra que ela se mantenha o **Modelo 2** (com uma consulta ao crédito) precisa de 0,999 de limiar de aprovação.

Por fim, a arquitetura provou-se robusta para classificação das propostas com o **Modelo 1** e o **Modelo 2**.

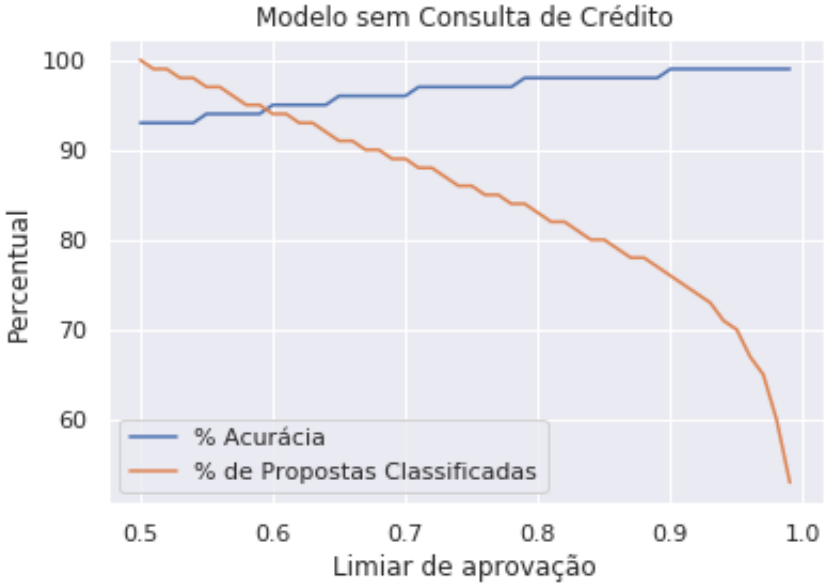

Figura 5. Percentual de acurácia e total de classificações para 96.041 propostas do Modelo 1.

Acurácia	Limiar M1	Limiar M2	**Modelo 1**	**Modelo 2**
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 4. Valores dos limiares necessários em cada modelo para satisfazer a acurácia exigida e os percentuais totais de classificação.

5. Conclusões

Construímos uma abordagem alternativa à análise automática de crédito mais eficiente que a análise atual e menos custosa. Utilizando uma técnica de Aprendizagem de Máquina, o algoritmo *Random Forest Regressor*, desenvolveu-se dois modelos preditivos com base no histórico de propostas, capazes de prever a situação aprovada ou reprovada de cada proposta.

Iniciamos a análise exploratória dos dados carregando um dataset de quase 200.000 propostas desde janeiro de 2018 até abril de 2019. A partir daí, removeu-se os *outliers*, os casos incoerentes, removeu-se atributos com menor representatividade estatística, removeu-se atributos que possuíam alta correlação entre si e, por fim, dividimos o dataset tratado em duas grandes partes de quase 100.000 propostas cada, escolhidas aleatoriamente, para treinarmos e testarmos os resultados.

Para construir os modelos, o primeiro esforço se concentrou em identificar um número ideal de árvores no parâmetro do *Random Forest Regressor*. Para a calibração do modelo, considerou-se uma sub-amostra de 30.000 propostas e gerou-se gráficos de

scores pela quantidade de árvores. Com isso, foi definido 100 árvores para o **Modelo 1** e 75 árvores para os outros modelos.

Foram implementadas duas metodologias capazes de avaliar a qualidade do modelo proposto. A primeira mediu a acurácia dos resultados sob de um limiar de certeza esperado. A segunda mediu o volume de propostas que o modelo foi capaz classificar sob o limiar de aprovação esperado. Avaliou-se a taxa de acurácia e a taxa do volume de contratos classificados por limiar de aprovação esperado. Foi visto que, quanto maior o limiar maior a acurácia sob a classificação 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. Se tomarmos, por exemplo, a acurácia de 97% o limiar de 0,75 do Modelo M1 classifica 86,56% das propostas e o Modelo M2 mais 4,04% do restante das propostas não classificadas.

No atual modelo 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 atuais são executadas por um sistema independente ao qual é submetida a proposta do cliente. Infelizmente, não é gravado log do resultado de todas as regras para cada proposta submetida, apenas para a regra que aprovou, reprovou ou repassou a proposta é gravado log.

Diante desse cenário no qual as regras não puderam ser reconstituídas completamente nem os resultados das avaliações das regras puderam ser aproveitados, precisamos 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.

O atual modelo utilizado na empresa classifica automaticamente aproximadamente 52% das propostas. Para a acurácia mais alta de 99% o modelo proposto já classificaria 69,19%, um ganho substancial. Assim, há uma redução considerável de propostas para análise manual por operadores de crédito, o que possui alto potencial de redução de custos com pessoal. Além disso, 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 consultarão o crédito do cliente enquanto o modelo atual consulta o crédito em cerca de 80% das propostas. Constatamos que a consulta ao crédito realizada hoje em muitos casos é desnecessária, uma vez que o modelo sem consulta conseguiu replicar muito bem as classificações, o que significa que a decisão de aprovar ou reprovar não tem tanto peso na informação sobre classificação de crédito do cliente quanto se imagina hoje.

Caberá à empresa responder qual o limiar aceitável para o modelo operar.

Importante ressaltar que nós não consideramos neste trabalho avaliar a qualidade

do crédito concedido, mas apenas implementamos um modelo que represente bem o atual comportamento da empresa ao avaliar uma proposta de crédito. Isso significa que o modelo 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.

Referências

Batista, G. E. d. A. P. A. (2003). *Pré-processamento de dados em aprendizado de máquina supervisionado*. Doutorado em ciências, Universidade de São Paulo, São Carlos.

de Oliveira Araújo, F. and Ribeiro, A. P. A. (2003). *Mercado de crédito brasileiro*. Organização Internacional do Trabalho, 1ª edition.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923.

Gertler, M. and Karadi, P. (2015). Monetary policy surprises, credit costs, and economic activity. *American Economic Journal: Macroeconomics*, 7(1):44–76.

Guyon, I. and Elisseeff, A. (2006). An introduction to feature extraction. In *Feature extraction*, pages 1–25. Springer.

Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE.

Kealhofer, S. (2003). Quantifying credit risk I: default prediction. *Financial Analysts Journal*, 59(1):30–44.

Rodrigues, D. S., Brasil, A. R. A., Costa, M. B., Komati, K. S., and Pinto, L. A. (2018). Uma análise comparativa de um algoritmo de aprendizado supervisionado para solicitações de empréstimo em uma plataforma peer-to-peer. In *Anais do XIV Simpósio Brasileiro de Sistemas de Informação*, pages 332–325, Porto Alegre, RS, Brasil. SBC.

UOL (2018). 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.

Vasconcellos, M. S. d. (2002). *Proposta de método para análise de concessões de crédito a pessoas físicas*. PhD thesis, Universidade de São Paulo.

Zogbi, P. and Castro, M. (2019). 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.