FUNDAÇÃO ESCOLA DE COMÉRCIO ÁLVARES PENTEADO

FECAP

CENTRO UNIVERSITÁRIO ÁLVARES PENTEADO

MESTRADO PROFISSIONAL EM ADMINISTRAÇÃO

JANE SIMÕES DE CASTRO

ESTUDO COMPARATIVO ENTRE METODOLOGIAS DE APRENDIZADO DE MÁQUINA E HÍBRIDAS APLICADAS A RISCO DE CRÉDITO

São Paulo

2019

JANE SIMÕES DE CASTRO

ESTUDO COMPARATIVO ENTRE METODOLOGIAS DE APRENDIZADO DE MÁQUINA E HÍBRIDAS APLICADAS A RISCO DE CRÉDITO

Artigo apresentado à Fundação Escola de Comércio Álvares Penteado - FECAP, como parte dos requisitos para obtenção de título de Mestre em Administração

Orientador: Prof. Dr. Joelson Oliveira Sampaio
Co-Orientador: Prof. Dr. Vinicius Augusto Brunassi Silva

São Paulo

2019

FICHA CATALOGRÁFICA

C355e

Castro, Jane Simões de

Estudo comparativo entre metodologias de aprendizado de máquina e híbridas aplicadas a risco de crédito / Jane Simões de Castro. - - São Paulo, 2019.

25 f.

Orientador: Prof. Dr. Joelson Oliveira Sampaio

Co-Orientador: Prof. Dr. Vinicius Augusto Brunassi Silva

Artigo (mestrado) – Fundação Escola de Comércio Álvares Penteado - FECAP - Centro Universitário Álvares Penteado – Programa de Mestrado Profissional em Administração com Ênfase em Finanças.

1. Administração de risco. 2. Inteligência artificial. 3. Aprendizado por computador. 4. Administração de crédito.

CDD 658.155

Bibliotecário responsável: Elba Lopes, CRB- 8/9622

JANE SIMÕES DE CASTRO

ESTUDO COMPARATIVO ENTRE METODOLOGIAS DE APRENDIZADO DE MÁQUINA E HÍBRIDAS APLICADAS A RISCO DE CRÉDITO

Artigo apresentado à Fundação Escola de Comércio Álvares Penteado - FECAP, como parte dos requisitos para a obtenção do título de Mestre em Administração.

COMISSÃO JULGADORA:

_____

Prof. Dr. Humberto Gallucci Netto
UNIFESP

_____

Prof. Dr. Vinícius Augusto Brunassi Silva
Fundação Escola de Comércio Álvares Penteado – FECAP

_____

Prof. Dr. Joelson Oliveira Sampaio
Fundação Escola de Comércio Álvares Penteado – FECAP
Professor Orientador – Presidente da Banca Examinadora

São Paulo, 10 de dezembro de 2019.

Estudo Comparativo Entre Metodologias De Aprendizado De Máquina E Híbridas Aplicadas A Risco De Crédito

Jane Simões De Castro
Mestre em Administração
E-mail: castro.jane@edu.fecap.br

Resumo

Para bancos e empresas que possuem operação de crédito, deter relações com clientes de alto risco aumenta a chance de inadimplência, a necessidade de alocação de capital e a exposição a prejuízos financeiros. Dessa forma, há interesse em aprimorar as avaliações de risco de crédito; e o cenário atual de Big Data fomenta o interesse em metodologias de inteligência artificial, uma vez que a assertividade dessas cresce à medida em que se aumenta a volumetria de dados utilizados. Essa dissertação tem por objetivo comparar metodologias quantitativas aplicáveis à gestão de risco de crédito e concluir se técnicas baseadas em inteligência artificial apresentam performance superior às técnicas tradicionais. Foram estudadas as metodologias Regressão Logística, Support Vector Machine, Random Forest, Gradient Boosting e Modelos híbridos, em visão pessoa física e visão pessoa jurídica. Para ambas visões, a comparação dos modelos via métricas de performance AUC, KS e Taxa de acerto indicou o Gradient Boosting como metodologia campeã.

Palavras-chaves: Risco de Crédito. Inteligência Artificial. Aprendizado de Máquina. SVM. Boosting. Modelos Híbridos.

Abstract

For banks and companies that offer credit operations, having relationships with high-risk customers increases the probability of default, the need for capital allocation and the exposure to financial losses. Thus, there is interest in improving credit risk evaluations; and the current Big Data scenario enhances the interest in artificial intelligence methodologies, since their accuracy increases as the volume of data also increases. This dissertation aims to compare quantitative methodologies applicable to credit risk management and to conclude whether the techniques based on artificial intelligence present better performance than traditional techniques. The study includes the Logistic Regression, Support Vector Machine, Random Forest, Gradient Boosting and Hybrid Models methodologies, for both individuals and enterprises. The study concludes that Gradient Boosting is the champion methodology in the comparation made through the performance metrics AUC, KS and Hit rate.

Keywords: Credit Risk. Artificial Intelligence. Machine Learning. SVM. Boosting. Hybrid Models.

1 Introdução

A sustentabilidade do ciclo de crédito de bancos e quaisquer outras empresas que operem com risco de crédito requer que ambos envolvidos na operação, tanto cedente quanto tomador, sejam fiéis às suas obrigações contratuais.

Relações com clientes mal selecionados geram maior chance de inadimplência e consequentes prejuízos financeiros. Em oposição, o resultado de expandir a carteira de clientes sem aumento da inadimplência é expresso em aumento de receita. Logo, é evidente que há interesse no desenvolvimento de novas técnicas que sejam capazes de gerar escores de crédito mais assertivos (Huang, Chen, & Wang, 2007).

Atualmente, a mensuração quantitativa do risco de crédito praticada pelas empresas, principalmente bancos, é comumente feita por metodologias clássicas de padrões lineares, com destaque para a regressão logística (Y. S. Chen & Cheng, 2013). Entretanto novas técnicas vêm ganhando notoriedade, sobretudo aquelas que operam com base em inteligência artificial (IA), como as de Machine Learning (aprendizado de máquina).

A assertividade dos algoritmos de IA tende a ser maior à medida em que aumenta-se a volumetria de dados utilizados. Logo, em um cenário como o atual, de crescente disponibilidade de dados e larga capacidade de armazenamento e processamento, pesquisas sobre Machine Learning passam a ser alvo de maior interesse (Burrell, 2016).

Em linha com este contexto, o objetivo desse trabalho é analisar se a utilização de metodologias de aprendizado de máquina no processo de modelagem de risco de crédito traz melhorias em relação a metodologias tradicionais.

Para alcançar esse objetivo e simultaneamente apresentar extensões em relação a trabalhos anteriores sobre o tema no cenário nacional, serão considerados, além de metodologias de Machine Learning isoladas, modelos híbridos propostos na literatura.

Para a geração dos resultados, serão desenvolvidos modelos de regressão logística como referência e os demais modelos como desafiantes. Os efeitos das técnicas selecionadas serão analisados tanto para pessoas físicas quanto para pessoas jurídicas, em âmbito nacional.

Ainda que a gestão de riscos eficiente envolva parâmetros políticos e morais que vão além da identificação de bons pagadores, como evidenciado pela crise do subprime de 2008, quando concedeu-se crédito imobiliário a uma vastidão de clientes de alto risco, visando o lucro excessivo a curto prazo (Ackermann, 2008), ter um processo decisório que selecione clientes adequados é elementar para garantir a perenidade do ciclo de crédito.

2 Revisão de literatura

Os escores resultante de modelos de riscos de crédito são insumo de diversos processos decisórios do cedente, como aceitar ou rejeitar solicitações de crédito, aumentar ou não limites de cartões de crédito já concedidos, efetuar cobranças com abordagem amigável ou agressiva, aprovar ou declinar renegociações de dívida inadimplentes, etc.

Como citado na seção anterior, os modelos de crédito mais usuais no mercado são análises multivariadas lineares, mas a inteligência artificial tem ganhado destaque devido a metodologias de aprendizado de máquina.

Tais metodologias possuem algumas subcategoriais. Uma delas são os Ensemble Methods, que utilizam as saídas de diversos modelos conjuntamente, a fim de melhorar a acurácia do escore final, como explicado no trabalho de Kwon, Choi e Suh (2013).

Similares aos métodos Ensemble, os Hybrid Models utilizam múltiplos resultados simultaneamente para incrementar a performance do modelo, porém enquanto os métodos Ensemble trabalham com a junção de vários modelos homogêneos fracos para criar um forte, os métodos Hybrid utilizam modelos heterogêneos combinados (Kazienko, Lughofer, & Trawiński, 2013).

Outra categoria, Deep Learning, engloba algoritmos mais sofisticados, baseados em redes neurais, e surgiu como resposta à necessidade de processar extensos volumes de dados (Zhang, Tan, Han, & Zhu, 2017).

A seguir serão citados trabalhos comparativos entre metodologias clássicas e de Machine Learning, cujas descrições técnicas serão apresentadas na seção seguinte.

Addo, Guegan e Hassani (2018) realizaram um estudo com dados de pessoas jurídicas de um banco privado europeu, comparando seis técnicas de Machine Learning, sendo quatro delas baseadas em Deep Learning. Como resultado, as performances dos modelos Deep Learning foram inferiores às performances obtidas com uso das metodologias Gradient Boosting e Random Forest.

No trabalho de comparação de metodologias de Ghatasheh (2014) com uso de dados cadastrais, demográficos e de histórico de crédito de banco alemães e australianos, o algoritmo de Random Forest apresentou resultados superiores a outras quinze técnicas. Além da melhor performance, os autores defenderam a técnica pela sua maior simplicidade e interpretabilidade em relação às outras testadas.

Para casos de metodologia híbrida, Kwon, Choi e Suh (2013) concluíram que a combinação das metodologias Bagging e Boosting apresentaram melhores resultados em

comparação a outras. O estudo de Chen, Ma e Ma (2009) concluiu que o uso híbrido de modelos de Support Vector Machine (SVM) com árvores de decisão e SVM com regressão por Splines apresentam performances superiores ao uso das mesmas técnicas isoladamente.

Chen et al. (2012) propõe um processo de modelagem híbrido de clusterização por K-means e SVM, para dados de cartão de crédito de um banco chinês. Concluem que para determinados pontos de corte, a classificação de bons e maus clientes é superior a outras técnicas.

No cenário brasileiro, Forti (2018) realizou uma comparação entre metodologias para modelos de cobrança, concluindo que, para esta etapa do ciclo de crédito, as técnicas de aprendizado de máquina apresentam melhores resultados em relação à regressão logística. Aniceto (2016) realizou uma comparação análoga, com modelos de risco de crédito, considerando a inadimplência dos clientes pessoas físicas já aprovados para uma linha de crédito de um banco brasileiro.

Na visão pessoa jurídica, Gregório (2018) analisou o desempenho da metodologia KMV, da Moody's, em relação a outros de Machine Learning. O KMV é uma metodologia de cálculo de probabilidade de descumprimento para corporações, baseado em dados de balanço, governança corporativa e macroeconômicos. Em seu trabalho, concluiu que esta técnica não apresenta ganho em comparação a técnica de Machine Learning XGBoost para as corporações brasileiras consideradas, analogamente a estudos feitos para corporações americanas.

A metodologia XGBoost também apresentou performance superior a ténicas de regressão logística e random forest no estudo de Marra (2019) para a previsão de dificuldades financeiras de empresas da américa Latina, considerando dados oriundos do Economática sobre empresas ativas ou canceladas.

Nos trabalhos citados, as métricas utilizadas para comparação de desempenho se repetem, sendo as mais utilizadas: area under the curve (AUC), root-mean-square-error (RMSE) e taxa de acurácia ou taxa de acerto. No ambiente corporativo, as medidas mais comuns para avaliação de modelos de risco de crédito são kolmogorov-smirnov (KS) e coeficiente de Gini (Rezac & Rezac, 2011).

Em sua conclusão, Addo et al. (2018) fazem uma ponderação relevante: apesar dos modelos baseado em IA trazerem uma expectativa de melhor performance, quão mais sofisticada a metodologia utilizada, menor a transparência do passo a passo do algoritmo e mais complexa a interpretabilidade do impacto de cada variável no escore calculado.

Tais considerações, juntamente com a avaliação da capacidade implantação de modelos e as exigências dos órgãos reguladores, devem sempre ser consideradas durante a escolha da metodologia a ser usada no desenvolvimento de escores de crédito.

3 Metodologia

Como mencionado, a técnica mais consolidada para modelagem de risco de crédito é a regressão logística (Chen & Cheng, 2013), cuja estrutura é uma equação linear, garantindo alta interpretabilidade e simples implantação.

Para os fins comparativos que são objetivo desse projeto, serão desenvolvidos modelos de regressão logística como referencial e modelos desafiantes com metodologias de Machine Learning, apresentadas a seguir, assim como modelos híbridos, que são compostos por combinações de metodologias.

Para mensurar e comparar a qualidade dos modelos propostos serão avaliadas as métricas mais presentes na literatura e no mercado, descritas na seção 4.2.

3.1 Metodologias ensemble

Como citado na revisão da literatura, os métodos Ensemble referem-se a uma combinação de modelos individualmente fracos que conjuntamente resultam em um modelo forte.

Esses métodos podem ainda ser classificados nas categorias Bagging e Boosting. O algoritmo de metodologias Bagging consiste na geração de subamostras dos dados originais, construção de modelos independentes para cada uma das subamostras e combinação dos modelos para um resultado único (Breiman, 1996).

As metodologias Boosting (Freund, Schapire, & Hill, 1996) também consistem em uma combinação de modelos, porém estes são dependentes e sequenciais, em que cada modelo objetiva minimizar o erro do modelo anterior.

3.1.1 Bagging

A metodologia mais disseminada desta categoria é a Random Forest, uma combinação aditiva de vários modelos estimados por árvores de decisão.

Para uma população de N indivíduos com K variáveis preditoras, estima-se múltiplas árvores independentes, de forma que cada uma considere uma amostra aleatória dos N indivíduos e uma amostra aleatória das K variáveis preditoras (Biau, Devroye, & Lugosi, 2008).

No contexto de risco de crédito, cada árvore retorna uma probabilidade estimada do indivíduo ser inadimplente, e a probabilidade final estimada pela Random Forest é uma combinação com pesos iguais destas probabilidades, como a média simples das probabilidades.

3.1.2 Boosting

O algoritmo mais conhecido desta categoria é o Adaboost, publicado por Freund et al. (1996).

Por esta metodologia estima-se um modelo inicial a partir de uma amostra aleatória do conjunto de dados. Em seguida, aplica-se este modelo em todo o conjunto de dados, atribui-se pesos maiores para as observações que obtiveram os maiores erros e gera-se nova amostra aleatória para seguir o processo. Com esta dinâmica, as observações com pior ajuste serão mais prováveis de estar na amostragem para a construção do modelo seguinte, cuja estimação deve explicá-las de forma mais acurada, diminuindo o erro final da combinação dos modelos.

Derivado do Adaboost, a metodologia Gradient Boosting (GB) usualmente é aplicada como a combinação de árvores de decisão. Após a estimação da árvore inicial, as sequenciais são desenvolvidas de modo a minimizar o resíduo da anterior, o que pode ser realizado por diferentes funções de perda, sendo mínimos quadrados a mais usual (Friedman, 2002).

Os processos de ajuste de modelos de GB devem considerar o teste de diferentes hiperparâmetros, tais quais quantidade máxima de árvores, profundidade máxima da árvore e número mínimo de indivíduos por nó, etc.

3.2 Support Vector Machine

SVM é um algoritmo de identificação de padrões e classificação de variáveis binárias, apresentado por Cortes e Vapnik (1995). A essência da metodologia é a otimização matemática para identificação de um hiperplano que maximize a separação das observações em dois perfis, distinguindo perfis binários de interesse (Bellotti & Crook, 2009).

O ajuste de SVM pode ser feito assumindo ou não a linearidade do hiperplano. Para casos de não-linearidade, é necessário analisar e decidir a melhor função de matemática.

3.3 Modelos híbridos

O estudo de Chen et al. (2009) propõe uma modelagem em duas etapas. Primeiramente estima-se uma árvore de decisão para selecionar as variáveis com maior poder discriminatório. Em seguida, estima-se um modelo por SVM apenas com as variáveis resultantes das primeira etapa.

O estudo de Chen et al. (2012) também sugere uma modelagem em duas etapas, porém a primeira etapa é utilizada para tratamento da variável resposta.

Inicialmente, realiza-se uma análise de cluster e classifica-se os grupos resultantes como taxa de inadimplência alta, média ou baixa. Então, constrói-se uma nova variável reposta, em que todas as observações pertencentes a clusters com alta taxa de inadimplência são marcadas como inadimplentes, todas as observações pertencentes a clusters com baixa taxa de inadimplência são marcadas como adimplentes, e todas as observações pertencentes a clusters com média taxa de inadimplência mantém a marcação original. Em seguida, estima-se um modelo por SVM com a nova variável resposta construída.

Espera-se que a primeira etapa auxilie na criação de grupos homogêneos que removam o impacto de observações atípicas, melhorando a performance do modelo final.

Essas duas abordagens serão testadas neste projeto, estendendo a segunda etapa para outras metodologias além de SVM.

3.4 Métricas

3.4.1 Taxa de acerto

Uma das métricas mais usuais nos trabalhos revisados é a taxa de acerto. Para calculá-la admite-se um valor *t*, tal que se $P(y_i = 1 / x_i) < t$, o cliente *i* é classificado como adimplente e caso contrário é classificado como inadimplente. Calcula-se então a taxa de acerto do modelo (Bellotti & Crook, 2009).

Apesar do cálculo simples, há um caráter subjetivo nessa métrica, ao passo que não existe um valor *t* padrão.

Para o presente projeto, será usado o corte t = 0,5. Ou seja, serão nomeados como inadimplentes os casos que apresentarem probabilidade de inadimplência maior que a probabilidade de adimplência, e vice-versa.

3.4.2 Curva ROC e AUC

O modelo pode apresentar dois erros: atribuir uma alta probabilidade de inadimplência a um bom cliente (erro do tipo 1) e atribuir uma baixa probabilidade de inadimplência a um mau cliente (erro do tipo 2). Altas e baixas probabilidades são relativas a um patamar *t*, como explicado no item anterior, e para cada patamar, é possível calcular as probabilidades de os erros acontecerem.

As probabilidades dos erros 1 e 2 acontecerem são respectivamente representados por α e β, e os valores de (1 – α) e (1 – β) são respectivamente chamados de especificidade e sensitividade. (Addo et al., 2018).

Para construção da curva ROC calcula-se a especificidade e sensitividade para diversos patamares *t,* gerando a representação gráfica abaixo.

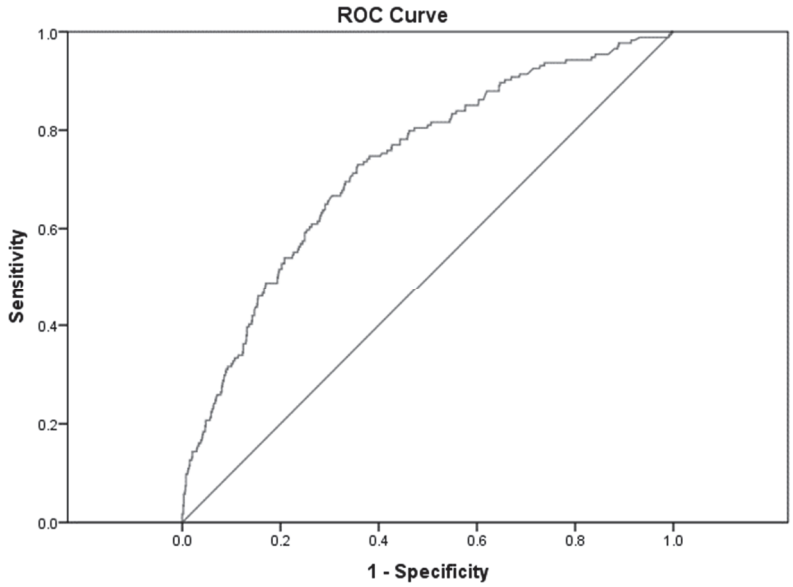

Figura 1. Curva ROC.

Qualquer ponto na curva diagonal representa um cenário em que a proporção de casos classificados corretamente é igual a proporção de casos classificados incorretamente, ou seja, o cenário esperado por uma classificação aleatória. Pontos acima dessa curva representam cenários em que a taxa de classificados corretamente é maior que a taxa de classificados incorretamente.

O índice AUC (area under the curve) representa a área abaixo da curva ROC, assumindo valores de 0 a 1, sendo que quão maior o seu valor, maior é a capacidade do modelo discriminar pagadores bons e maus.

3.4.3 KS

Para a estatística Kolmogorov-Smirnov (KS) calcula-se a distribuição acumulada de bons e de maus por escore. O valor KS é a distância máxima entre essas distribuições para o mesmo escore, representado graficamente abaixo (Rezac & Rezac, 2011).

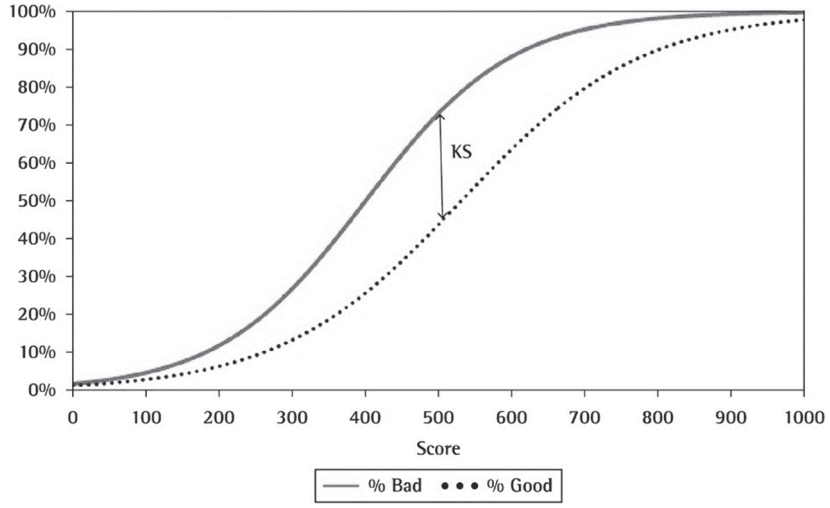

Figura 2. KS

A estatística KS assume valores de 0 a 100%, sendo que quão maior o seu valor, maior é a capacidade do modelo discriminar clientes bons e maus.

4 Dados

Para o desenvolvimento deste projeto foram utilizadas informações provenientes da Serasa Experian. Por ser um bureau com dados de empresas e indivíduos de todo o Brasil, o uso desses dados não apresenta o viés de outros trabalhos que utilizaram bancos como fonte de dados: ao utilizar dados de crédito bancários provenientes das próprias instituições, trabalha-se apenas com os clientes aprovados pelas mesmas, ou seja, há uma pré-seleção de melhores clientes.

Na visão pessoa jurídica há um ganho adicional em relação à literatura: trabalhar com uma visão pessoa jurídica mais granular do que as mapeadas nos estudos nacionais apresentados na seção 2, pois serão consideradas empresas ativas de todos os portes e segmentos, bem como seus dados presentes na Serasa, e não apenas empresas com informações públicas.

Foram utilizadas amostras mensais, de pessoas físicas e de pessoas jurídicas, com dados nas referências de janeiro a dezembro de 2017. Essas são observadas por 12 meses, como é evidenciado no item 4.1, de forma que foram utilizadas informações até dezembro de 2018, data mais recente autorizada pela empresa.

O público de pessoas jurídicas considera apenas empresas ativas na receita federal, e o de pessoas físicas apenas maiores de 18 anos de idade, uma vez que públicos diferentes desses não são elegíveis a crédito.

As amostras possuem volumetria de 20 mil por mês e tipo de pessoa, e foram construídas de forma que um mesmo indivíduo ou empresa não estivesse presente em mais de um mês, evitando um possível viés.

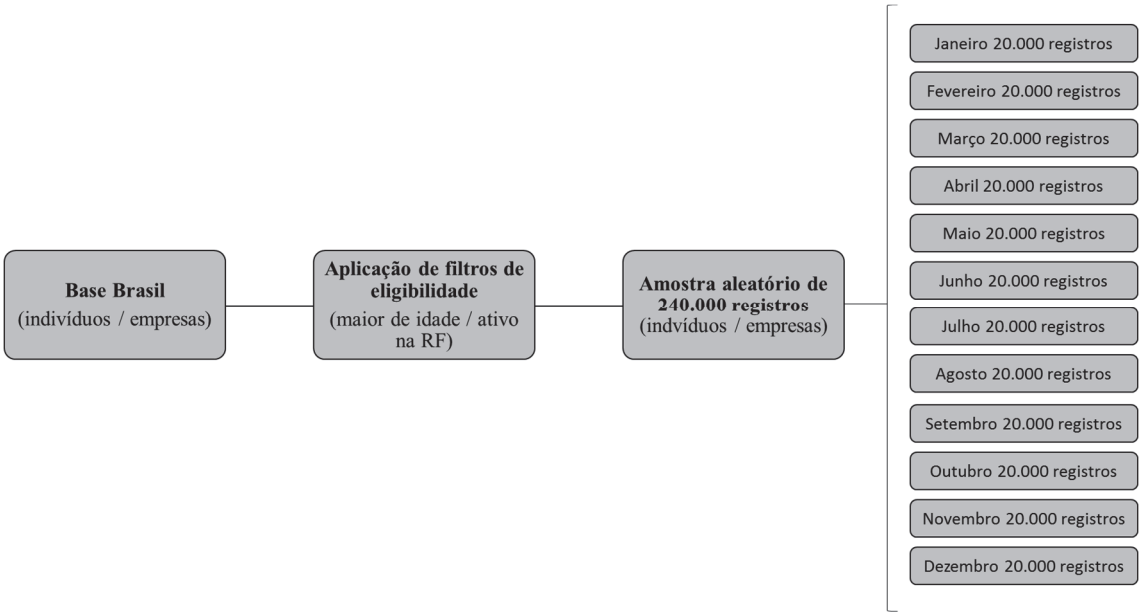

Figura 3. Processo de amostragem.

4.1 Variável resposta

O objetivo de modelos de risco de crédito é prever a chance de um cliente se tornar inadimplente após um determinado período de tempo, logo, é necessário definir o conceito de inadimplência a ser utilizado.

Como o banco de dados possui o histórico de dívidas de crédito vencidas e não pagas, essa informação foi utilizada de forma que se considerou inadimplente o cliente que possua ao menos uma dívida vencida e não paga um ano após a data de referência da amostra a que pertence. Caso contrário, marcou-se o cliente como adimplente.

Consequentemente, o modelo estima a probabilidade de um cliente se tornar inadimplente durante os 12 meses seguintes à data de referência. Foram devidamente removidos da amostra os casos de indivíduos e empresas que já inadimplentes na data de referência, para não haver viés nos resultados.

Seguem abaixo os gráficos das taxas da variável resposta, ou seja, das taxas de clientes inadimplentes na amostra.

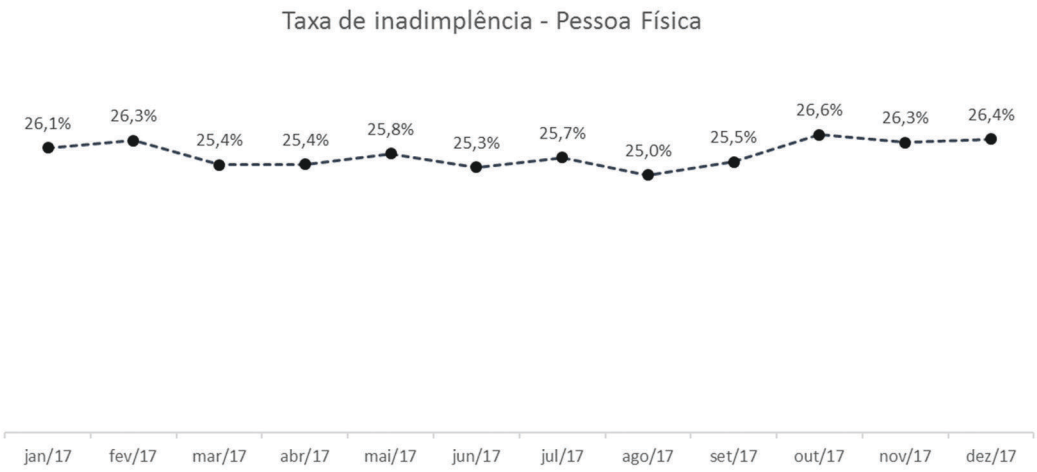

Figura 4. Variável resposta do público pessoa física

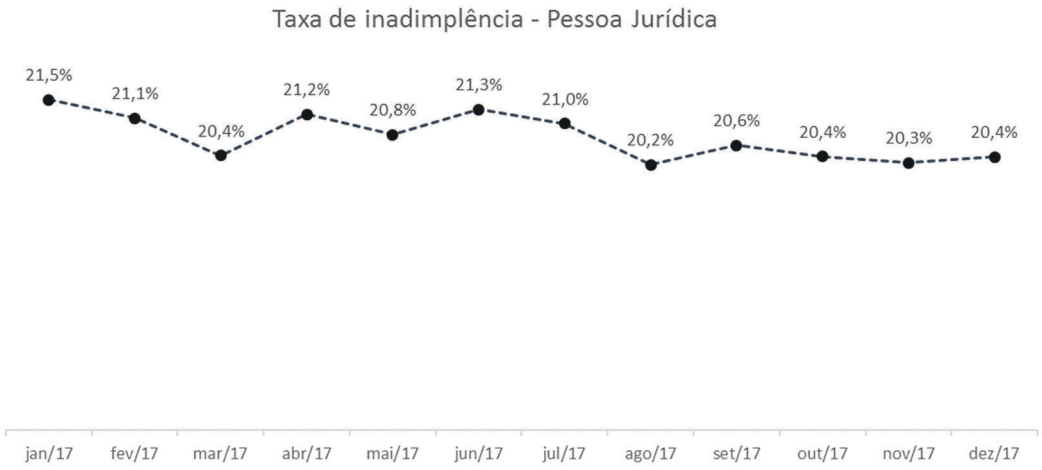

Figura 5. Variável resposta do público pessoa jurídica.

4.2 Variáveis preditoras

As demais variáveis do banco de dados foram consideradas como variáveis preditoras.

Para o público pessoa física havia inicialmente 700 variáveis, das quais 419 foram descartadas por baixo preenchimento ou por baixa variabilidade (percentil 5% igual ao percentil 95%). Assim, para a modelagem, foram consideradas 281 variáveis, das seguintes categorias:

Tabela 1
Variáveis preditoras do público pessoa física

Categoria	Quantidade de variáveis
Cadastral	3
Endereço e telefone	3
Renda	7
Demográficas	16
Empresário	10
Histórico de consultas	109
Histórico de dívidas	133

Analogamente, para o público pessoa jurídica, foram descartadas 560 das 1.212 iniciais, restando para a modelagem, 652 variáveis.

Tabela 2
Variáveis preditoras do público pessoa jurídica

Categoria	Quantidade de variáveis
Cadastral	12
Pagamentos	22
Histórico de consultas	90
Histórico de consultas - Sócios PF	130
Histórico de consultas - Todos Sócios	136
Histórico de dívidas	73
Histórico de dívidas - Sócios PF	85
Histórico de dívidas - Todos sócios	104

4.3 Amostra de validação

Uma preocupação durante o desenvolvimento de modelos é o overfitting: superestimação dos dados, de modo que o modelo apenas performe para uma amostra específica.

Para mitigar esse problema, uma boa prática é a separação da base em amostra de desenvolvimento e amostra de validação: a amostra de desenvolvimento é utilizada para realizar o ajuste do modelo e a amostra de validação, para analisar se o modelo desenvolvido mantém sua performance em um período diferente do qual ele foi treinado.

Para tal finalidade, foram considerados para o desenvolvimento os meses de janeiro a setembro/2017, e para validação, os demais.

5 Desenvolvimento dos modelos

Nos tópicos seguintes estão pontuadas particularidades de cada metodologia durante o desenvolvimento dos modelos e os resultados dos mesmos, mensurados pelas métricas previamente apontadas.

5.1 Regressão logística

Como critério para estimação de regressão logística foram selecionadas variáveis que não apresentassem multicolinearidade (VIF < 10) e indicassem alta significância (p-valor < 0,01%). As variáveis foram utilizadas de acordo com a própria natureza, isto é, variáveis contínuas não passaram por processos de categorização.

As equações estimadas para previsão de inadimplência de pessoas físicas e pessoas jurídicas estão representadas abaixo.

$P(inadimplência\ PF_i)$ = -10,14 + 0,01 * quantidade de consultas a crédito nos últimos 5 anos $_i$ – 0,10 * quantidade de consultas a crédito bancário no último 1 ano $_i$ – 0,37 * quantidade de credores com os quais já possuiu dívida $_i$ – 0,03 * quantidade de vezes em que possuiu dívida $_i$ – 0,23* quantidade de dívidas no último ano $_i$ – 0,01 * valor de dívidas em empresas de telecomunicações nos últimos 5 anos $_i$ + e $_i$

$P(inadimplência\ PJ_i)$ = -1,12 – 0,17 * quantidade de dívidas nos últimos 2 anos $_i$ – 0,26 * quantidade de consultas por indústrias nos últimos 30 dias $_i$ – 0,90 * quantidade de consultas por bancos nos últimos 90 dias $_i$ – 0,11 * quantidade de dívidas dos sócios $_i$ - 0,04 * valor de dívidas nos últimos 3 anos $_i$ + 0,01 * tempo desde a última consulta de bancos + e $_i$

Essas equações resultaram nas métricas de performance abaixo, que serão usadas como base para comparação com os demais modelos.

Tabela 3
Resultados da metodologia Regressão Logística

	Taxa de acerto	KS	AUC	Gini
		Desenvolvimento		
PF	80,8%	45,8%	80,0%	59,9%
PJ	82,1%	42,3%	76,7%	53,4%
		Validação		
PF	80,1%	45,2%	79,9%	59,8%
PJ	82,3%	41,5%	75,8%	51,7%
		Razão		
PF	99%	99%	100%	100%
PJ	100%	98%	99%	97%

5.2 Random Forest

Para a modelagem de Random Forest, inicialmente foram gerados modelos utilizando possíveis combinações entre número de árvores e número de variáveis. Os erros quadráticos médios para essas combinações estão representados abaixo, para PF e PJ.

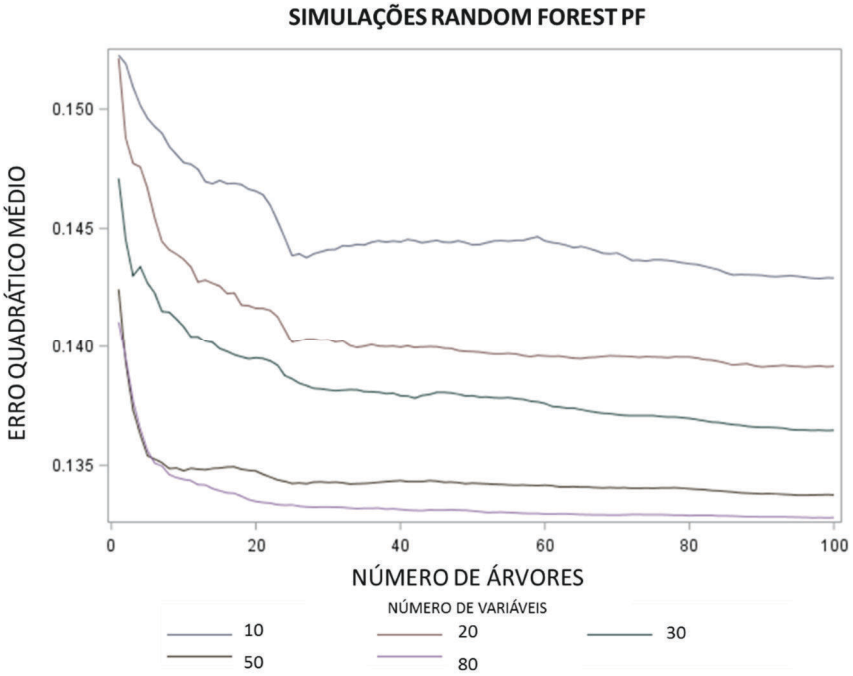

Figura 6. Simulações da metodologia Random Forest PF.

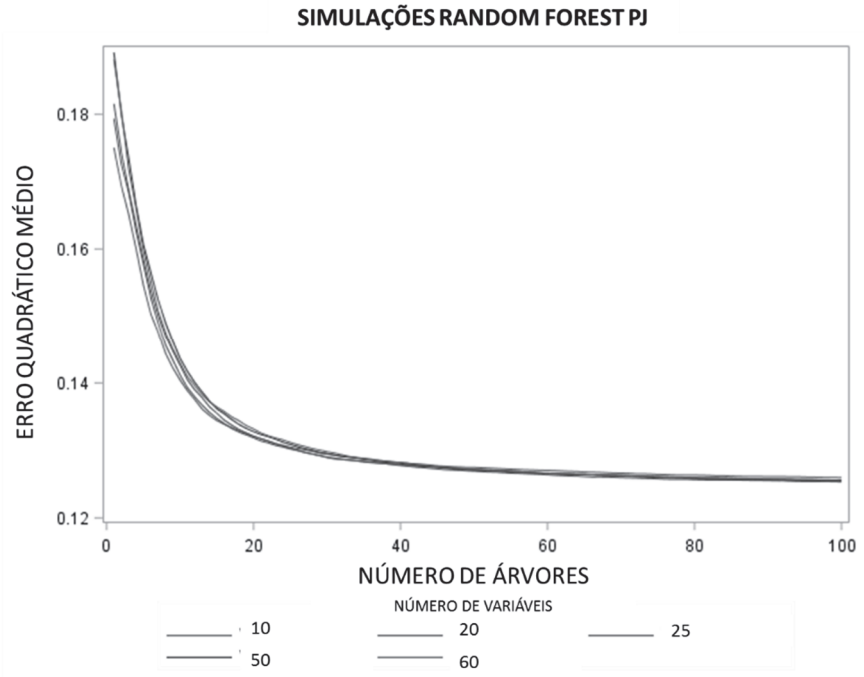

Figura 7. Simulações da metodologia Random Forest PJ.

O objetivo desses gráficos é a identificação da combinação de menor número de árvores e menor número de variáveis que apresente um erro quadrático médio próximo ao mínimo possível.

Considerando esses pontos, para PF identificou-se que seriam adequados o mínimo de 80 variáveis e máximo de 40 árvores; e para PJ, máximo de 10 variáveis e máximo de 50 árvores. Entretanto, essas combinações resultaram em overfitting, com perda relevante de performance na amostra de validação.

Ao passo que quanto mais árvores e variáveis maior é a chance do overfitting ocorrer, foram realizados diversos testes diminuindo os parâmetros, até encontrar resultados satisfatórios.

As combinações mais adequadas foram 25 árvores e 35 variáveis para PF, e 30 árvores e 50 variáveis para PJ. Tais modelos resultaram em erros quadráticos médios próximo ao mínimo possível de cada público (0,139 e 0,135, respectivamente), e nos indicadores de performance abaixo:

Tabela 4
Resultados da metodologia Random Forest

	Taxa de acerto	KS	AUC	Gini
		Desenvolvimento		
PF	80,1%	46,1%	80,5%	60,9%
PJ	82,5%	40,3%	76,7%	53,3%
		Validação		
PF	79,4%	45,5%	80,2%	60,4%
PJ	82,5%	40,3%	76,7%	53,3%
		Razão		
PF	99%	99%	100%	99%
PJ	100%	100%	100%	100%

5.3 Gradient Boosting

Diferentemente da modelagem por Random Forest, o algoritmo para modelagem de Gradient Boosting calcula intrinsicamente o número ótimo de árvores.

Para a amostra PF foram estimadas 298 árvores, e para a amostra PJ, 148. Tais conjuntos de modelo resultaram nos seguintes indicadores de performance:

Tabela 5
Resultados da metodologia Gradient Boosting

	Taxa de acerto	KS	AUC	Gini
		Desenvolvimento		
PF	82,1%	49,2%	82,4%	64,8%
PJ	83,0%	46,4%	80,5%	61,0%
		Validação		
PF	81,2%	48,0%	81,8%	63,6%
PJ	83,2%	45,6%	79,6%	59,2%
		Razão		
PF	99%	98%	99%	98%
PJ	100%	98%	99%	97%

5.3 Support Vector Machine

Para a modelagem com SVM é fundamental a escolha da melhor função de kernel, que são conjuntos de funções matemáticas para delinear o hiperplano. Nesse projeto foram testadas as funções linear, radial e polinomial de 2, 3 e 4 graus.

Tanto para PF quando para PJ os resultados de todas as funções foram muito similares, com diferença apenas na quarta casa decimal. Logo, foi selecionada a função mais simples, linear, que gerou os seguintes resultados:

Tabela 6
Resultados da metodologia SVM

	Taxa de acerto	KS	AUC	Gini
		Desenvolvimento		
PF	78,5%	35,4%	72,9%	45,8%
PJ	82,3%	44,2%	79,2%	58,5%
		Validação		
PF	77,9%	35,4%	72,8%	45,5%
PJ	82,4%	42,1%	77,7%	55,5%
		Razão		
PF	99%	100%	100%	99%
PJ	100%	95%	98%	95%

5.4 Modelos híbridos por árvore de decisão

Na proposta de Chen, Ma e Ma (2009), utiliza-se a metodologia de árvore de decisão em uma etapa inicial de seleção de variáveis, e então aplica-se uma segunda metodologia para estimação das probabilidades, considerando apenas as variáveis selecionadas na primeira etapa.

Por essa primeira etapa espera-se limitar o universo de variáveis apenas às mais relevantes e construir um modelo performático utilizando um número reduzido de parâmetros. Para ambientes corporativos, esse ganho é de extrema importância, uma vez que traz redução de tempo e custos de implantação.

Pela árvore de decisão na etapa inicial foram selecionadas 39 variáveis de um total de 281 para a amostra PF, e 33 variáveis de 352 para a amostra PJ. Considerando apenas essas variáveis, foram desenvolvidos modelos por Random Forest, GB e SVM, cujos resultados foram:

Tabela 7
Resultados da metodologia híbrida árvore de decisão + Random Forest

	Taxa de acerto	KS	AUC	Gini
		Desenvolvimento		
PF	80,0%	45,8%	80,2%	60,4%
PJ	83,0%	43,1%	78,3%	56,6%
		Validação		
PF	79,5%	45,1%	80,0%	59,9%
PJ	83,0%	43,1%	78,3%	56,6%
		Razão		
PF	99%	99%	100%	99%
PJ	100%	100%	100%	100%

Tabela 8
Resultados da metodologia híbrida árvore de decisão + Gradient Boosting

	Taxa de acerto	KS	AUC	Gini
		Desenvolvimento		
PF	81,7%	48,3%	81,6%	63,2%
PJ	83,2%	46,6%	80,8%	61,6%
		Validação		
PF	81,2%	47,3%	81,5%	62,9%
PJ	83,2%	45,5%	79,9%	59,8%
		Razão		
PF	99%	98%	100%	100%
PJ	100%	97%	99%	97%

Tabela 9
Resultados da metodologia híbrida árvore de decisão + SVM

	Taxa de acerto	KS	AUC	Gini
		Desenvolvimento		
PF	80,7%	48,0%	80,9%	61,9%
PJ	81,7%	43,0%	76,9%	53,7%
		Validação		
PF	79,9%	47,9%	80,9%	61,8%
PJ	82,0%	41,8%	75,9%	51,8%
		Razão		
PF	99%	100%	100%	100%
PJ	100%	97%	99%	96%

5.5 Modelos híbridos por cluster

Para a proposta de Chen et al. (2012), foram aplicadas inicialmente análises de cluster.

Para PF, os dados foram classificados em 5 clusters com as seguintes características:

Tabela 10
Resultados da análise de cluster para público pessoa física

Cluster	% Amostra	% Inadimplentes
1	31,7%	27,5%
2	51,7%	13,1%
3	2,4%	79,7%
4	14,1%	59,0%
5	0,2%	100,0%

Tomando como base que a taxa de inadimplência da carteira é 25,8%, inferiu-se que o cluster 1 representa um perfil de inadimplência dentro da média, o cluster 2, um perfil de baixa inadimplência e os demais, um perfil de alta inadimplência.

Logo, pela metodologia sugerida, para as modelagens na segunda etapa o público do cluster 1 foi mantido com sua variável resposta original, o público do cluster 2 considerado como adimplente e os demais como inadimplentes.

Analogamente para PJ, foram estimados 3 cluster:

Tabela 11
Resultados da análise de cluster para público pessoa jurídica

Cluster	% Amostra	% Inadimplentes
1	0,5%	100,0%
2	94,6%	18,7%
3	4,9%	59,9%

Apesar do público PJ se mostrar mais homogêneo, foi possível aplicar o mesmo raciocínio da PF, de forma que os públicos dos clusters 1 e 3 foram considerados inadimplentes e do cluster 2, adimplentes.

Considerando essa nova variável resposta, foram desenvolvidos modelos por Random Forest, GB e SVM, com os seguintes resultados:

Tabela 12
Resultados da metodologia híbrida cluster + Random Forest

	Taxa de acerto	KS	AUC	Gini
		Desenvolvimento		
PF	74,4%	46,1%	80,5%	60,9%
PJ	82,8%	42,3%	77,7%	55,3%
		Validação		
PF	73,6%	45,5%	80,2%	60,4%
PJ	83,0%	43,1%	78,3%	56,6%
		Razão		
PF	99%	99%	100%	99%
PJ	100%	100%	100%	100%

Tabela 13
Resultados da metodologia híbrida cluster + Gradient Boosting

	Taxa de acerto	KS	AUC	Gini
		Desenvolvimento		
PF	81,6%	48,2%	82,1%	64,3%
PJ	80,1%	33,9%	72,3%	44,6%
		Validação		
PF	80,6%	47,5%	81,6%	63,2%
PJ	80,3%	33,2%	71,8%	43,6%
		Razão		
PF	99%	99%	99%	98%
PJ	100%	98%	99%	98%

Tabela 14
Resultados da metodologia híbrida cluster + SVM

	Taxa de acerto	KS	AUC	Gini
		Desenvolvimento		
PF	78,1%	34,0%	71,5%	43,1%
PJ	82,4%	45,4%	78,8%	57,5%
		Validação		
PF	77,2%	33,7%	71,4%	42,7%
PJ	82,5%	42,8%	77,1%	54,1%
		Razão		
PF	99%	99%	100%	99%
PJ	100%	94%	98%	94%

6 Resultados

Nessa seção estão reapresentados os resultados de todos os modelos desenvolvidos, a fim de facilitar a comparação.

Tabela 15
Comparação de resultados para público pessoa física

		Desenvolvimento				Validação			
		Taxa de acerto	KS	AUC	Gini	Taxa de acerto	KS	AUC	Gini
	Regressão Logística	80,8%	45,8%	80,0%	59,9%	80,1%	45,2%	79,9%	59,8%
	SVM	78,5%	35,4%	72,9%	45,8%	77,9%	35,4%	72,8%	45,5%
Metodologias Ensemble	RF	80,1%	46,1%	80,5%	60,9%	79,4%	45,5%	80,2%	60,4%
	GBM	82,1%	49,2%	82,4%	64,8%	81,2%	48,0%	81,8%	63,6%
	SVM	80,7%	48,0%	80,9%	61,9%	79,9%	47,9%	80,9%	61,8%
Metodologias híbrida - Árvore	RF	80,0%	45,8%	80,2%	60,4%	79,5%	45,1%	80,0%	59,9%
	GBM	81,7%	48,3%	81,6%	63,2%	81,2%	47,3%	81,5%	62,9%
	SVM	78,1%	34,0%	71,5%	43,1%	77,2%	33,7%	71,4%	42,7%
Metodologias híbrida - Cluster	RF	74,4%	46,1%	80,5%	60,9%	73,6%	45,5%	80,2%	60,4%
	GBM	81,6%	48,2%	82,1%	64,3%	80,6%	47,5%	81,6%	63,2%

Tabela 16
Diferença de resultados em relação à regressão logística para público pessoa física

		Desenvolvimento				Validação			
		Taxa de acerto	KS	AUC	Gini	Taxa de acerto	KS	AUC	Gini
	SVM	-2,4%	-10,3%	-7,1%	-14,1%	-2,2%	-9,8%	-7,1%	-14,3%
Metodologias Ensemble	RF	-0,7%	0,4%	0,5%	1,0%	-1,4%	-0,2%	0,2%	0,5%
	GBM	1,3%	3,5%	2,5%	4,9%	0,4%	2,3%	1,8%	3,7%
	SVM	-0,1%	2,3%	1,0%	1,9%	-0,9%	2,2%	0,9%	1,9%
Metodologias híbrida - Árvore	RF	-0,9%	0,0%	0,2%	0,5%	-1,3%	-0,6%	0,0%	0,0%
	GBM	0,9%	2,5%	1,6%	3,3%	0,4%	1,6%	1,5%	3,0%
	SVM	-2,7%	-11,8%	-8,4%	-16,9%	-3,6%	-12,1%	-8,6%	-17,2%
Metodologias híbrida - Cluster	RF	-6,4%	0,4%	0,5%	1,0%	-7,2%	-0,3%	0,2%	0,5%
	GBM	0,8%	2,5%	2,2%	4,3%	-0,2%	1,8%	1,7%	3,3%

Tabela 17
Comparação de resultados para público pessoa jurídica

		Desenvolvimento				Validação			
		Taxa de acerto	KS	AUC	Gini	Taxa de acerto	KS	AUC	Gini
	Regressão Logística	82,1%	42,3%	76,7%	53,4%	82,3%	41,5%	75,8%	51,7%
	SVM	82,3%	44,2%	79,2%	58,5%	82,4%	42,1%	77,7%	55,5%
Metodologias Ensemble	RF	82,5%	40,3%	76,7%	53,3%	82,5%	40,3%	76,7%	53,3%
	GBM	83,0%	46,4%	80,5%	61,0%	83,2%	45,6%	79,6%	59,2%
	SVM	81,7%	43,0%	76,9%	53,7%	82,0%	41,8%	75,9%	51,8%
Metodologias híbrida - Árvore	RF	83,0%	43,1%	78,3%	56,6%	83,0%	43,1%	78,3%	56,6%
	GBM	83,2%	46,6%	80,8%	61,6%	83,2%	45,5%	79,9%	59,8%
	SVM	82,4%	45,4%	78,8%	57,5%	82,5%	42,8%	77,1%	54,1%
Metodologias híbrida - Cluster	RF	82,8%	42,3%	77,7%	55,3%	82,8%	42,3%	77,7%	55,3%
	GBM	80,1%	33,9%	72,3%	44,6%	80,3%	33,2%	71,8%	43,6%

Tabela 18
Diferença de resultados em relação à regressão logística para público pessoa jurídica

		Desenvolvimento				Validação			
		Taxa de acerto	KS	AUC	Gini	Taxa de acerto	KS	AUC	Gini
	SVM	0,2%	1,9%	2,5%	5,1%	0,0%	0,6%	1,9%	3,8%
Metodologias Ensemble	RF	0,5%	-2,0%	0,0%	-0,1%	0,5%	-2,0%	0,0%	-0,1%
	GBM	0,9%	4,1%	3,8%	7,6%	1,1%	3,3%	2,9%	5,8%
	SVM	-0,4%	0,7%	0,1%	0,3%	-0,1%	-0,5%	-0,8%	-1,6%
Metodologias híbrida - Árvore	RF	0,9%	0,8%	1,6%	3,2%	0,9%	0,8%	1,6%	3,2%
	GBM	1,1%	4,3%	4,1%	8,2%	1,1%	3,2%	3,2%	6,4%
	SVM	0,3%	3,1%	2,1%	4,1%	0,5%	0,5%	0,4%	0,7%
Metodologias híbrida - Cluster	RF	0,7%	-0,1%	1,0%	1,9%	0,7%	-0,1%	1,0%	1,9%
	GBM	-2,0%	-8,4%	-4,4%	-8,8%	-1,8%	-9,2%	-4,9%	-9,9%

O modelo com melhor performance para pessoas físicas foi o desenvolvido por Gradient Boosting. Para pessoas jurídicas, o modelo híbrido com seleção de variáveis por árvore de decisão e segunda etapa por GB foi o campeão, mas com pouca superioridade em relação ao GB puro.

Em oposição, os resultados dos modelos por SVM se mostraram sempre inferiores a todas as metodologias, inclusive à regressão logística.

Outro ponto interessante foram os modelos com metodologia híbrida de cluster resultando em baixa performance, indicando que a alteração da variável resposta não foi relevante no processo de modelagem.

7 Conclusão

As metodologias baseadas em Gradient Boosting apresentaram métricas de performance superiores a outras metodologias. Tal conclusão está em linha com os resultados de trabalhos nacionais citados na revisão metodológica, mesmo este projeto se diferenciando em qualidade do público, por não apresentar viés de seleção bancária, e na inclusão de metodologias híbridas.

O ganho de performance nos modelos de crédito implicam diretamente em diminuição de perda financeira por inadimplência em bancos e financeiras. Entretanto, mesmo ao apresentar superioridade, os modelos de machine learning enfrentam barreiras de entradas nas instituições pela falta de conhecimento dos algoritmos das metodologias.

Apesar da não-linearidade da metodologia incorrer em maior dificuldade de interpretabilidade, também possibilita menor rigidez e maior rapidez no processo de modelagem, uma vez que etapas de seleção de variáveis não são essenciais.

Dessa forma, a insegurança gerada pela falta de conhecimento da fórmula do modelo pode ser mitigada pela capacidade de se atualizar o modelo com maior frequência, mantendo a operação de crédito controlada e suportadas por modelos desenvolvidos com metodologia de performance superior.

Como sugestão de próximos passos, indica-se o estudo expandido com uso de variáveis oriundas do cadastro positivo, que entra em vigor no Brasil a partir desse ano.

Referências

Ackermann, J. (2008). The subprime crisis and its consequences. *Journal of Financial Stability*, *4*(4), 329–337.

Addo, P. M., Guegan, D., & Hassani, B. (2018). Credit risk analysis using Machine and Deep Learning models. *Risks*, *6*, 38–57.

Aniceto, M. C. (2016). *Estudo comparativo entre técnicas de aprendizado de máquina para estimação de risco de crédito* (Dissertação de Mestrado). Universidade de Brasília, Brásilia, DF, Brasil.

Bellotti, T., & Crook, J. (2009). Support vector machines for credit scoring and discovery of significant features. *Expert Systems with Applications*, *36*(2), 3302–3308..

Breiman, L. E. O. (1996). Bagging Predictors. *Machine Learning, 24*(2), 123–140.

Burrell, J. (2016, June). How the machine “thinks:” Understanding opacity in Machine Learning algorithms. *Big Data & Society*, 1–12. doi: 10.1177/2053951715622512

Chen, W., Ma, C., & Ma, L. (2009). Mining the customer credit using hybrid support vector machine technique. *Expert Systems with Applications*, *36*(4), 7611–7616.

Chen, W., Xiang, G., Liu, Y., & Wang, K. (2012). *Credit risk Evaluation by hybrid data mining technique*. *3*, 194–200. https://doi.org/10.1016/j.sepro.2011.10.029

Chen, Y. S., & Cheng, C. H. (2013). Hybrid models based on rough set classifiers for setting credit rating decision rules in the global banking industry. *Knowledge-Based Systems*, *39*, 224–239.

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning, 20*(5), 273–

297.

Forti, M. (2018). *Técnicas de Machine Learning aplicadas na recuperação de crédito do mercado brasileiro* (Dissertação de Mestrado). Fundação Getúlio Vargas, São Paulo, SP, Brasil.

Freund, Y., Schapire, R. E., & Hill, M. (1996). Experiments with a New Boosting Algorithm. *Machine Learning: Proceedings of the Thirteenth International Conference, 7(5),* 148-156.

Friedman, J. H. (2002) Stochastic gradient boosting. C*omputational Statistics and Data Analysis, 28*, 367-378.

Gerard, B., Devroye, L., & Lugosi, G. (2008). Consistency of Random Forests and other averaging classifiers. *Journal of Machine Learning Research*, *9* (8), 2015–2033.

Ghatasheh, N. (2014, November). Business analytics using random forest trees for credit risk prediction: A comparison study. *International Journal of Advanced Science and Technology*, *72*, 19–30.

Gregório, R. L. (2018). *Modelo híbrido de avaliação de risco de crédito para corporações brasileiras com base em algoritmos de aprendizado de máquina* (Dissertação de Mestrado). Universidade Católica de Brasília, Brasília, DF, Brasil

Huang, C. L., Chen, M. C., & Wang, C. J. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications*, *33*(4), 847–856.

Kazienko, P., Lughofer, E., & Trawiński, B. (2013). Hybrid and ensemble methods in machine learning. *Journal of Universal Computer Science*, *19*(4), 457–461.

Kwon, J., Choi, K., & Suh, Y. (2013). Double ensemble approaches to predicting firms' credit rating. *Association for Information Systems, 13(1).* 158-163.

Rezac, M., & Rezac, F. (2011). How to measure the quality of credit scoring models. *Finance a Uver - Czech Journal of Economics and Finance*, *61*(5), 486–507.

Vinícius Nogueira Marra. (2019). *Previsão de dificuldades financeiras em empresas latino americanas via aprendizagem de máquina* (Dissertação de Mestrado). Universidade Federal de Uberlândia, Uberlândia, MG, Brasil.

Zhang, L., Tan, J., Han, D., & Zhu, H. (2017). From machine learning to deep learning: progress in machine intelligence for rational drug discovery. *Drug Discovery Today*, *22*(11), 1680–1685.