Uso de Auto Machine Learning na Modelagem de Loss Given Default: Desafios e Oportunidades

Douglas Beserra Pinheiro*
GCB Investimentos e FGV/SP

RESUMO

O uso de Machine Learning (ML) difundido em diversas áreas ainda encontra barreiras no mercado financeiro devido à complexidade de seus algoritmos, interpretabilidade de seus resultados e por limitações regulatórios, por isso são chamados de modelos "caixa-preta". Esse artigo explora um algoritmo de Auto Machine Learning (AutoML), uma evolução dos modelos de ML, na estimação de Loss Given Default (LGD) de uma carteira de crédito. Os resultados mostraram que: o AutoML é muito superior às metodologias tradicionais (regressão linear e árvore de decisão) tanto em termos acurácia quanto ordenação; o tamanho da amostra é muito importante para a performance do modelo; o tempo de execução do algoritmo não apresentou diferença relevante nos cenários avaliados; as variáveis relevantes foram em sua maioria as mesmas selecionadas pela metodologia tradicional. A recente evolução das ferramentas de AutoML e a consistência dos resultados indica que essa metodologia pode ser aplicada nos demais parâmetros de perda e nos mais diversos produtos de crédito, reduzindo o tempo de desenvolvimento, aumentando a qualidade dos resultados e atendendo os requisitos do novo Acordo de Basiléia.

Palavras-chave: *Crédito, Avaliação de Modelos, Machine Learning*

Classificação JEL: H81, C52, C61

ABSTRACT

The widespread use of Machine Learning (ML) in several areas still faces barriers in the financial market due to the complexity of their algorithms, the interpretability of their results and regulatory limitations, which is why they are called "black box" models. This article explores an Auto Machine Learning (AutoML) algorithm, an evolution of ML models, in estimating the Loss Given Default (LGD) of a credit portfolio. The results showed that: AutoML is far superior to traditional methodologies (linear regression and decision tree) both in terms of accuracy and ordering; sample size is very important for model performance; the algorithm execution time did not show any relevant difference in the situations evaluated; the relevant variables were mostly the same as those selected by the traditional methodology. The recent evolution of AutoML tools and the consistency of results indicate that this methodology can be applied to other loss parameters and to the most diverse credit products, thereby reducing development time, increasing the quality of results, and meeting the requirements of the New Basel Accord.

Key Words: *Credit, Model Assessment, Machine Learning*

*GCB Investimentos, São Paulo, Brasil. Email: douglas.pinheiro@gcbinvestimentos.com e FGV/SP

1. Introdução

Desde 1980 o uso de técnicas de Aprendizado de Máquina Supervisionado, ou Supervised Machine Learning (doravante ML), em áreas como reconhecimento de fala, imagem, escrita, veículos autônomos, previsão não linear de séries temporais (Breiman, 2001; Israel *et al.*, 2020) e mais recentemente previsões no mercado financeiro relacionadas a risco de crédito, sinistro de seguros e fraudes (Tuggener *et al*., 2019) onde modelos paramétricos tinham pouca utilidade e a performance era muito relevante tornou-se muito difundida. EBA (2021) define de forma suscinta ML como o "...*campo da ciência da computação que lida com o desenvolvimento de modelos cujos parâmetros são estimados automaticamente a partir dos dados, com pouca ou nenhuma intervenção humana*". Simplificadamente, é possível replicar em larga escala o que seria o aprendizado humano, de tal forma que o algoritmo realize previsões acuradas a partir dos padrões detectados e refinados sobre os dados observados (Guégan e Hassani, 2018).

O sucesso nessas aplicações e o aumento massificado do volume de dados incentivaram grandes empresas a desenvolver e disponibilizar recursos de computação e armazenamento em nuvem[1], aumentando o poder computacional e reduzindo seu custo, o que viabilizou o uso de modelos complexos no mercado financeiro sem acarretar um elevado custo mesmo para as novas empresas de crédito (Bazarbash, 2019). Alonso e Carbó (2020), por exemplo, observaram um ganho de até 20% na performance[2] de modelos de inadimplência desenvolvidos por ML em comparação com modelos tradicionais e destacam que eles são usados principalmente no crédito massificado, onde há uma maior abundância e qualidade de dados necessários para o treinamento dos modelos. EBA (2021) também destaca a limitação dos modelos de regressão linear no variado e complexo universo do *Big Data*.

Apesar da extensa literatura sobre a aplicação de ML em modelos de previsão de inadimplência, pouco ainda foi estudado sobre os demais parâmetros de perda esperada. Desde a publicação do Acordo de Basileia II (BIS, 2004), as instituições financeiras em todo o mundo podem optar por estimar a perda esperada utilizando modelos internos.

[1] Atualmente as três principais soluções são a AWS da Amazon, o Google Cloud Platform e a Azure da Microsoft.

[2] Aqui especificamente em termos de AUC, indicador que varia de 0 (pior ordenação) a 1 (melhor ordenação), destacando que a melhor ordenação significa uma melhor separação entre tomadores de crédito bons e ruins.

Dessa forma a perda esperada decorrente do risco de crédito pode ser obtida pelo produto dos seguintes componentes sobre a exposição[3]:

- **PD (*Probability of Default*):** indica a expectativa de ocorrência da inadimplência ou descumprimento do empréstimo;
- **EAD (*Exposure at Default*):** corresponde ao valor esperado do empréstimo no momento da inadimplência ou descumprimento;
- **LGD (*Loss Given Default*):** refere-se à perda econômica, calculada em função do EAD, decorrente da inadimplência ou descumprimento.

Importante destacar que para o uso do parâmetro estimado a instituição financeira precisa demonstrar ao órgão supervisor que as estimativas são realistas em relação ao perfil de risco dos empréstimos (Lima, 2008). Um modelo com uma boa performance é aquele que consegue prever com maior precisão o valor do parâmetro de interesse. Além do cálculo da perda esperada, o LGD também é utilizado no cálculo do capital regulatório das instituições financeiras, em modelos de apreçamento, recuperação de crédito, apuração de valor justo contábil, em testes de stress e no gerenciamento de risco (Scandizzo, 2016; EBA, 2021).

Apesar da evidente evolução dos métodos de ML, especialmente após a publicação de BIS II, o foco dos reguladores permanece limitado na identificação dos fatores explicativos dos modelos (EBA, 2021), em melhorar a comparabilidade das abordagens adotadas pelas instituições financeiras (Guégan e Hassani, 2018) e no aprofundamento da definição dos conceitos básicos, por exemplo inadimplência. Enquanto isso questões ligadas ao mundo da tecnologia avançada, entre eles ML e Inteligência Artificial (IA) permaneceram em segundo plano (EBA, 2021). Uma evolução recente ocorreu com desenvolvimento de algoritmos de Aprendizado Automático de Máquina (doravante AutoML) que procuram substituir a ação humana em todas as etapas de desenvolvimento e escolha de modelos. Este artigo explora um algoritmo específico e o compara com as metodologias tradicionais em termos de performance e consistência das estimativas de LGD.

Este artigo preenche uma importante lacuna na área de risco de crédito, sendo o primeiro a explorar um processo de AutoML na estimação do LGD e avaliando, além da

[3] Neste artigo, a exposição se refere ao saldo devedor do empréstimo no momento da estimação dos componentes de risco. Neste caso a perda esperada é estimada em termos nominais.

diferença da performance, os principais determinantes do resultado em termos de variáveis explicativas. Isso é importante pois é a principal crítica dos regulares ao uso de modelos chamados "caixa-preta", ou seja, onde não fica clara a contribuição das variáveis para as estimativas e, consequentemente, para a gestão do crédito.

Com relação à performance, o AutoML mostrou-se muito superior aos métodos tradicionais tanto em termos de acurácia quanto ordenação. O AutoML reduziu o erro em mais de 50% em todas as métricas avaliadas, sendo robusto até mesmo na menor amostra de treino. Considerando as métricas de ordenação, a diferença é ainda significativa, em alguns casos melhorando cerca de 10 vezes o indicador. Com relação ao tamanho da amostra, de fato a performance do AutoML cai consideravelmente com a sua redução, o que pode ser um fator limitante para algumas aplicações quando não há um número relevante de observações, mas mesmo nos cenários mais críticos a performance ainda foi muito superior à metodologia tradicional. Finalmente não foi observada uma variação relevante de performance do modelo quando o tempo de processamento foi reduzido de 3 horas para apenas 1 hora, o que é relevante pois permite o desenvolvimento de um maior número de simulações na mesma máquina sem perda relevante qualidade do modelo.

Apesar da utilização de mais de 100 variáveis em todos os cenários de treino do AutoML, a relevância medida pelo valor de Shap [4] indicou que as variáveis mais importantes foram em sua maioria as mesmas selecionadas pelo método tradicional, o que reforça a consistência e segurança dos resultados obtido, bem como a capacidade dos modelos de ML em capturar relações complexas entre as variáveis, otimizando o seu uso. Esse conjunto de resultados indicam uma vantagem clara do uso de AutoML em termos de performance, custo e uma menor necessidade de especialização dos envolvidos no processo, uma vez que a maioria das etapas do desenvolvimento são realizados pelo algoritmo. Evidentemente existe ainda um amplo campo de pesquisa na área de ML e AutoML em crédito e debater esses resultados e alternativas é a forma mais produtiva de aumentar o uso desses métodos, melhorando as estimativas e garantindo uma melhor gestão de risco e resultados das instituições financeiras.

Este artigo está organizado da seguinte forma. A Seção 2 detalha o LGD e suas características, bem como traz um histórico sobre a evolução da modelagem. A Seção 3

[4] Os valores de Shapley derivam de pesquisas desenvolvidas por Lloyd Shapley no contexto dos jogos cooperativos.

apresenta as hipóteses testadas e a metodologia empregada. A Seção 4 descreve os dados utilizados e respectivas variáveis. A Seção 5 apresenta e discute os resultados empíricos observados. Finalmente, a Seção 6 conclui este artigo e apresenta uma breve discussão sobre os desafios do uso de AutoML.

2. Modelagem de Loss Given Default

O primeiro desafio nos estudos de LGD refere-se à sua apuração, pois existem muitas abordagens possíveis para o seu cálculo (BIS 2005, Scandizzo, 2016). As principais são: 1) *Market LGD*, baseado na comparação dos preços de mercado de títulos inadimplentes ou empréstimos negociáveis logo após a inadimplência com seu valor de face; 2) *Implied Market* LGD, derivado do spread de crédito de títulos adimplentes pois refletem a perda esperada percebida pelo mercado; 3) *Historical* LGD, aplicável ao varejo, o LGD se baseia nas perdas observadas e nas estimativas de PD de longo prazo; e, 4) *Workout* LGD, baseado no conjunto de fluxos de caixa recuperados após a inadimplência resultantes do processo de cobrança, propriamente descontado dos custos operacionais e do custo do dinheiro no tempo.

Normalmente empréstimos bancários são mantidos no balanço das instituições até sua a liquidação ou eventual baixa por prejuízo, então a abordagem normalmente utilizada para estimar o LGD é a *Workout* e ela contempla três tipos de perdas (Lima, 2008; Silva, 2009): i) a perda do principal; ii) receitas de juros não recebidos e iii) a perda relacionada às despesas relativas à cobrança e recuperação do crédito. Se o fluxo de caixa entre a data da inadimplência e o final do processo de recuperação é conhecido, então o LGD pode ser apurado conforme fórmula a seguir para uma determinada operação:

$$LGD_i = 1 - \frac{\sum_i R_i(r) - \sum_i C_i(r)}{EAD_i} \qquad (1)$$

onde, R_i refere-se à recuperação (seja ela monetária ou não), C_i aos custos (diretos e indiretos) incorridos e r representa o fator de desconto, fundamental para expressar o fluxo de caixa em termos de valores no momento da inadimplência. Normalmente o *LGD* é observado no intervalo entre 0 e 1, onde 0 significa que toda a perda (neste caso o *EAD*) foi recuperada e 1 que não houve recuperação. Na prática, o LGD pode assumir valor maior que um quando, por exemplo, não há recuperação alguma após a inadimplência, mas existem custos relacionados às tentativas de cobrança. Por outro lado, se a

recuperação é maior que o EAD, quando o devedor, por exemplo, paga as taxas de mora e o custo de cobrança foi baixo, o LGD pode ser negativo.

Na literatura, os principais determinantes do LGD são classificados em:

1) **Garantias:** o tipo da garantia (se real ou não, por exemplo, uma casa ou carro), a senioridade da garantia na execução e a sua liquidez (Lima, 2008; Schuermann, 2004).
2) **Ciclo Econômico:** períodos de recessão, além de aumentarem a inadimplência, também diminuem a capacidade de regularização das dívidas e o valor das garantias. Estudos indicam uma persistente correlação entre PD e LGD (Lima, 2008).
3) **Indústria:** quando um setor econômico está em crise, além da queda na rentabilidade, normalmente são prejudicados também o acesso ao crédito e o valor dos ativos próprios utilizados como garantias (Qi e Zhou, 2011). Além disso, setores com mais ativos tangíveis possuem maior recuperação (Schuermann, 2004).
4) **Contraparte:** a capacidade financeira da contraparte, seu volume de receitas ou renda e seu endividamento são importantes tanto na determinação da PD quando do LGD. Isso se aplica tanto às empresas quanto às pessoas (Zhou et al., 2018).

Uma vez calculado e identificados os fatores que afetam o LGD, o desafio seguinte refere-se à escolha da melhor técnica estatística para a sua previsão. Modelos simples de regressão e árvores de decisão tendem a ser robustos e promovem um bom entendimento dos fatores que o influenciam, porém normalmente apresentam baixa performance em termos de acurácia (Schuermann, 2004; Loterman et al, 2012).

Segundo Breiman (2001), os dois principais objetivos da análise de dados são predição, ou seja, a habilidade de prever os valores utilizando variáveis explicativas e a extração de informação, neste caso entender como as variáveis explicativas afetam a variável explicada. Para atender esses objetivos são utilizadas duas abordagens, a primeira chamada de “Cultura de Modelagem de Dados” assume um modelo de dados estocástico onde os parâmetros são estimados a partir dos dados observados e são finalmente utilizados para previsão ou extração de informação (p.e. regressão linear, modelo de Cox, regressão logística etc.). A segunda abordagem chamada de “Cultura de Modelagem Algorítmica” não presume nenhuma forma funcional na relação entre os dados e os considera complexos e desconhecidos em sua natureza (p.e. redes neurais, árvore de decisão etc.). Na primeira abordagem percebe-se uma preocupação maior com a análise dos efeitos das variáveis enquanto na segunda a ênfase maior é na qualidade da previsão

do modelo. Neste artigo chamamos a primeira abordagem de Metodologia Tradicional e a segunda abordagem como Metodologia Avançada (ou similarmente Machine Learning citado inicialmente).

A árvore de decisão em sua forma mais simples, segundo Bazarbash (2019), é construída tomando 2 decisões simples: 1) qual característica (variável explicativa) dividir e, 2) qual o valor limite da estatística para a divisão. Outros fatores podem ser determinados, por exemplo, a quantidade mínima de observações em cada nó folha ou a quantidade máxima de níveis que ela possa assumir. Incluindo certas restrições que limitam a sua estrutura, a árvore pode ser facilmente interpretada e monitorada, daí a sua escolha frequente para modelagem de LGD em instituições financeiros (EBA, 2021). Dessa forma, tanto a árvore de decisão construída sob restrições que facilitem sua interpretação quanto a regressão linear simples são avaliadas e classificadas neste artigo como Metodologia Tradicional.

Apesar de já existirem muitas soluções em Machine Learning no momento da publicação do Acordo de Basiléia II, essa abordagem foi preterida tanto nas pesquisas acadêmicas quanto nas instituições financeiras devido aos requisitos de validação interna de modelos (BIS, 2005) que em especial demandava "*... a validação do modelo deve incluir, por exemplo, uma revisão qualitativa da técnica de construção do modelo estatístico, a relevância dos dados usados para construir o modelo considerando o segmento de negócios específico do banco, a forma se seleção dos principais fatores de risco e se eles são economicamente significativos"*. Dessa forma, para os modelos terem seu uso aprovado existia uma ênfase muito grande em justificar os parâmetros escolhidos e essa escolha em modelos lineares era baseada principalmente no efeito esperado e no p-valor. Mesmo considerando as limitações dessa abordagem (Greenland *et al.*, 2016; Breiman, 2001) era amplamente difundida e aceita.

Entre os estudos que aplicaram a metodologia tradicional para estimação do LGD, Zhou *et al*. (2018) exploraram a mesma fonte de dados da *Lending Club* utilizada nesse artigo. Essas informações são especialmente interessantes pois a empresa, fundada em 2007 nos Estados Unidos, foi pioneira na exploração da internet para viabilizar empréstimos *peer-to-peer*. Ela é responsável pela intermediação, avaliação de risco, operacionalização e cuida de todo o relacionamento com o investidor e o tomador do crédito, eliminando assim a necessidade de um banco no processo crédito. Utilizando regressão linear Zhou *et al*. (2018) observaram que o rating, o prazo do empréstimo,

tempo no emprego, condição de moradia, endividamento, histórico de atrasos nos últimos 2 anos e o salário foram fatores relevantes na determinação do LGD. O R^2 das regressões foi particularmente baixo e variou entre 5% e 5,8%.

Silva *et al*. (2009) avaliaram 9557 empréstimos entre 2003 e 2007 no Brasil e através do uso de regressão censurada (Tobit) constataram que o nível de atividade econômica, a existência de garantias associadas, o valor do empréstimo e a ocorrência de uma renegociação prévia influenciavam o comportamento da LGD. O R^2 ajustado das regressões variou entre 7% e 9%. Dermine *et al.* (2006) estudaram o LGD em uma amostra de pequenas e médias empresas no período de 1995 a 2005 em Portugal e, uma vez que observaram uma acentuada distribuição bi-modal optaram pelo uso da regressão logística. Os autores constataram que o valor do empréstimo, as garantias, o setor da indústria, tempo de existência da empresa e o ano da inadimplência foram relevantes estatisticamente na determinação do LGD. O pseudo R^2 apresentou valores baixos, entre 8% e 18%. Yashkir e Yashkir (2013) avaliaram vários modelos paramétricos em uma amostra de 4275 títulos acompanhados pela S&P que apresentaram default entre 1981 e 2011 e constataram que a escolha do modelo é menos relevante do que a disponibilidade de variáveis explicativas, e a melhor especificação apresentou um R^2 de 39%.

Na abordagem tradicional existe uma ênfase relevante em identificar os fatores que afetam o LGD e para a adequada performance do modelo espera-se que esses fatores apresentem o efeito esperado pela teoria. Se bem desenvolvido, a expectativa é que o modelo mantenha seu nível de performance por mais tempo e eventuais alterações nas variáveis explicativas indicariam mudanças do mercado que poderiam afetar o LGD, e assim serviriam de alerta para a eventual atualização do modelo. Isso é especialmente importante no LGD *Workout* de produtos com garantia, onde todo o ciclo de recuperação pode levar anos. Os trabalhos citados mostram que há uma boa consistência das variáveis selecionadas, porém a performance dos modelos representada pelo R^2 é extremamente baixa, o que mostra a dificuldade de modelos paramétricos em capturar a complexa relação entre as variáveis e o LGD (Qi e Zhao, 2011).

Importante destacar também que, apesar da necessidade de teste em amostra independente do treinamento (BIS, 2004; BIS, 2005) para mitigar o risco de *overfitting* e erro de especificação, nenhum dos artigos desenvolvidos na metodologia tradicional incluiu esse teste na avaliação da qualidade dos modelos apresentados, o que limita o alcance de suas conclusões. O *overfitting* refere-se à boa performance do modelo na

amostra de treinamento em relação a uma performance ruim na amostra de teste, mais frequentemente observado em modelos complexos (Shuermann, 2004).

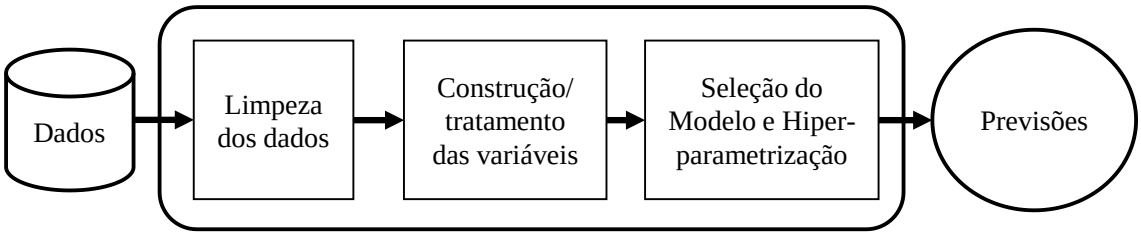
Loterman *et al*. (2012), foram pioneiros nas aplicações da Metodologia Avançada na estimação do LGD. Utilizando 6 bases de dados de bancos e produtos distintos avaliaram 6 técnicas de modelagem lineares e 4 não lineares, bem como uma combinação delas resultando em 24 processos distintos testados. As bases variavam de 3.351 a 119.211 observações e de 12 a 44 variáveis explicativas. Como o objetivo era a comparação das técnicas, o artigo não discutiu a respeito dos fatores que determinaram o LGD, mas em termos de performance o R^2 dos modelos variou de 4% a 43% (atingindo 13,8% na amostra de empréstimos pessoais similar ao estudado aqui), com clara vantagem observada nos modelos não paramétricos, em especial redes neurais e SVM (Support Vector Machine). Yao, Crook e Andreeva (2015) defendem que a superioridade de modelos não paramétricos decorre da não dependência de uma relação prévia entre a distribuição do LGD e as variáveis explicativas, tornando esses modelos muito mais flexíveis. Os autores testaram o SVR (Support Vector Regression) e algumas de suas variações contra outros algoritmos em uma amostra de 1413 títulos inadimplidos de grandes empresas americanas entre 1985 e 2012 reportados pela Moody's e obtiveram um R^2 de 70% com o modelo semi-paramétrico de mínimos quadrados SVR.

A mesma vantagem para modelos não paramétricos foi constatada por Qi e Zhao (2011) através de uma amostra de 3751 empréstimos bancários e títulos corporativos inadimplentes no período de 1985 a 2008 reportados pela Moody's. As melhores performances em termos de R^2 foram do modelo de redes neurais que alcançou 57,6% e de árvore de decisão com 82,7%, o que decorreu, segundo os autores, da maior capacidade desses modelos em capturar as relações não lineares entre as variáveis contínuas e o LGD. Papoušková e Hajek (2019) testaram modelos não paramétricos nos empréstimos da Lending Club do período de 2016 a 2018 em uma amostra de 82.095 empréstimos e constataram que o modelo de Random Forest em dois estágios (classificação e estimação) apresentou o melhor resultado. Eles obtiveram um erro quadrático médio (RMSE) de 0,053, um erro absoluto médio (MAE) de 0,026 e um R^2 de 0,934. Importante destacar que apresar de utilizar a metodologia de k-fold para o treino e teste dos modelos, Papoušková e Hajek (2019) não realizaram nenhum teste em amostra independente, ou seja, cujas observações não tenham sido utilizadas em nenhum treino.

Nos estudos de risco de crédito, apesar de os modelos de ML apresentarem uma performance superior aos modelos tradicionais, essa vantagem varia muito em função do algoritmo escolhido, da estrutura dos dados e da amostra utilizada no desenvolvimento do modelo (Bazarbash, 2019; Guegán e Hassani, 2018).

Zöller e Huber (2021) descrevem simplificadamente o desenvolvimento de um modelo de ML pelas seguintes etapas: 1) limpeza dos dados, onde são identificadas e tratadas observações com informações faltantes, *outliers*, formato inconsistente para o modelo e outros possíveis erros, 2) construção de novas variáveis ou combinação entre elas para enriquecer o modelo e, 3) desenvolvimento do modelo, onde a partir de um algoritmo escolhido seus hiperparâmetros são definidos para maximizar a performance das previsões. Os autores também destacam que, dependendo do algoritmo utilizado, a exclusão de variáveis irrelevantes ou redundantes pode ser realizada na etapa 2, ou seja, antes do refinamento do modelo e o processo pode ser repetido para diferentes algoritmos utilizados na etapa 3. Além disso, cabe destacar que o desenvolvimento de um modelo de Machine Learning adequado requer um vasto conhecimento dos algoritmos envolvidos e o seu desenvolvimento é suscetível a erros (Tuggener et al., 2019), tal como no desenvolvimento de qualquer modelo quando há interação humana. A figura 1 ilustra o fluxo apresentado.

Figura 1
Fluxo de Desenvolvimento (Pipeline) de um Modelo de ML



Adaptado de Zöller e Huber (2021).

No processo de modelagem tradicional os parâmetros são obtidos a partir do treinamento do modelo, por exemplo, temos os coeficientes de um modelo de regressão ou a segmentação de um nó em uma árvore de decisão que otimizam a performance do modelo. Os hiperparâmetros, ao contrário, são definidos no início do processo de aprendizado e em modelos de ML são extremamente relevantes na determinação da performance, como exemplos simples de hiperparâmetros temos em modelos de árvore de decisão a quantidade mínima de observações em determinado nó final ou a quantidade

máxima de níveis que determinamos para a árvore, já em modelos de regressão temos o p-valor mínimo que podemos determinar para a inclusão de uma variável no modelo. Em modelos de ML, a adequada escolha de hiperparâmetros é fundamental para a obtenção da melhor estimativa possível e, atualmente, existem vários algoritmos aplicados, entre eles Feurer e Hutter (2019) citam o *Grid Search*, onde o algoritmo percorre um grupo de hiperparâmetros pré-especificados, o *Random Serach*, em que o processo de escolha é aleatório e finalmente a Otimização Bayesiana, que é o método utilizado pelo algoritmo testado nesse artigo e detalhado na respectiva seção.

O Aprendizado Automático de Máquina (AutoML), uma evolução do ML tradicional, tem o objetivo de substituir todas as etapas do desenvolvimento de um modelo de Machine Learning, extremamente custosas em termos de consumo de tempo humano, por um algoritmo que procura otimizar, além da performance, o tempo e custo de processamento. Segundo Zöller e Huber (2021), soluções comerciais de AutoML foram desenvolvidas de tal forma que o profissional não precisa sequer de conhecimentos de programação para a sua utilização. Versões comerciais mais simples datam da década de 90 e eram focadas principalmente na etapa de hiper-parametrização automática.

Segundo Escalante (2020) um dos primeiros algoritmos completos de aprendizado foi o *Particle Swarm Model Selection* (PSMS), desenvolvido em 2006, e ele já incluía as etapas de pré-processamento de dados, seleção e extração de variáveis explicativas, classificação de modelos e otimização dos respectivos hiperparâmetros. Uma segunda onda de modelos, chamada de *Sequential Model-Based Optimization* (SMBO), foi desenvolvida a partir de 2010, e a sua ideia era avaliar uma série de possíveis modelos para estimar o desempenho e guiar a hiper-parametrização dos algoritmos, acelerando assim o processo de otimização pela redução do espaço de testes. Finalmente a onda mais recente, chamada de *Neural Architeture Search (NAS)*, passou a ser desenvolvida a partir de 2017 e foca na procura da melhor arquitetura e parametrização de modelos de aprendizado profundo (do inglês, *Deep Learning Models)*.

Zöller e Hubber (2021) avaliaram a performance de 6 algoritmos de AutoML, selecionados entre os mais populares do GitHub[5] e não constataram uma diferença significante de performance entre eles. O mesmo resultado foi obtido por Tuggener *et al*.

[5] Trata-se de uma das mais relevantes plataformas de armazenamento e controle de versão de códigos-fonte, tanto públicos quanto privados.

(2019) na avaliação de 4 algoritmos. O Auto-Sklearn[6] foi avaliado em ambos os estudos mencionados e ele foi selecionado neste artigo pois é oriundo da segunda onda de modelos e, portanto, além de apresentar uma solução completa para tratamento de dados e escolha dos algoritmos, foi vencedor de várias competições de performance de AutoML (Tuggener et al., 2019) e é o mais adotado atualmente (Escalante, 2020). Ele utiliza um *meta-learner* para direcionar a seleção do melhor modelo, além de testar uma combinação dos melhores modelos para o ajuste do modelo final.

O Auto-Sklearn foi desenvolvido no laboratório de inteligência Artificial da Universidade de Freiburg inicialmente em 2015 e construído sobre a estrutura do Scikit-learn[7], uma das mais relevantes bibliotecas de ciência de dados em linguagem Python. Sua segunda versão, detalhada em Feurer *et al*. (2019, 2022) foi a utilizada neste artigo e ela é capaz de estimar modelos de classificação binária, classificação multi-categórica e de regressão. Seu objetivo, bem como de qualquer outro algoritmo de AutoML, é definir como pré-processar os dados e encontrar o melhor algoritmo e seus hiperparâmetros para uma base de dados tabular qualquer. Seus principais componentes são:

1) **Dados.** Os dados brutos devem ser necessariamente numéricos e separados em amostras de aprendizado e teste;
2) **Meta-learning.** Uma série de 38 características (ou *meta-features*), tais como assimetria, curtose, número de classes, número de observações etc. de 140 conjuntos de dados de referência do OpenML[8] foram tabulados e treinados previamente usando o processo de otimização bayesiana explicada a seguir. Os hiperparâmetros que apresentaram os melhores resultados em cada conjunto de dados foram armazenados e servem como referência para novos conjuntos de dados. Assim, baseado na similaridade entre as características do novo conjunto de dados e as informações internalizadas no algoritmo medidos pela distância L1[9], são priorizados os hiperparâmetros dos 25 conjuntos de dados de menor distância, ou seja, aqueles que mais se assemelham ao novo conjunto de dados, reduzindo

[6] Detalhes podem ser obtidos em https://automl.github.io/auto-sklearn/master/.

[7] Detalhes disponíveis em https://scikit-learn.org/stable/.

[8] Detalhes disponíveis em www.openml.org, cujo repositório possui atualmente quase 5000 conjuntos de dados abertos.

[9] Na geometria de Manhattan, ou também chamada de métrica do táxi, a distância entre dois pontos é obtida pela soma das diferenças absolutas de suas coordenadas.

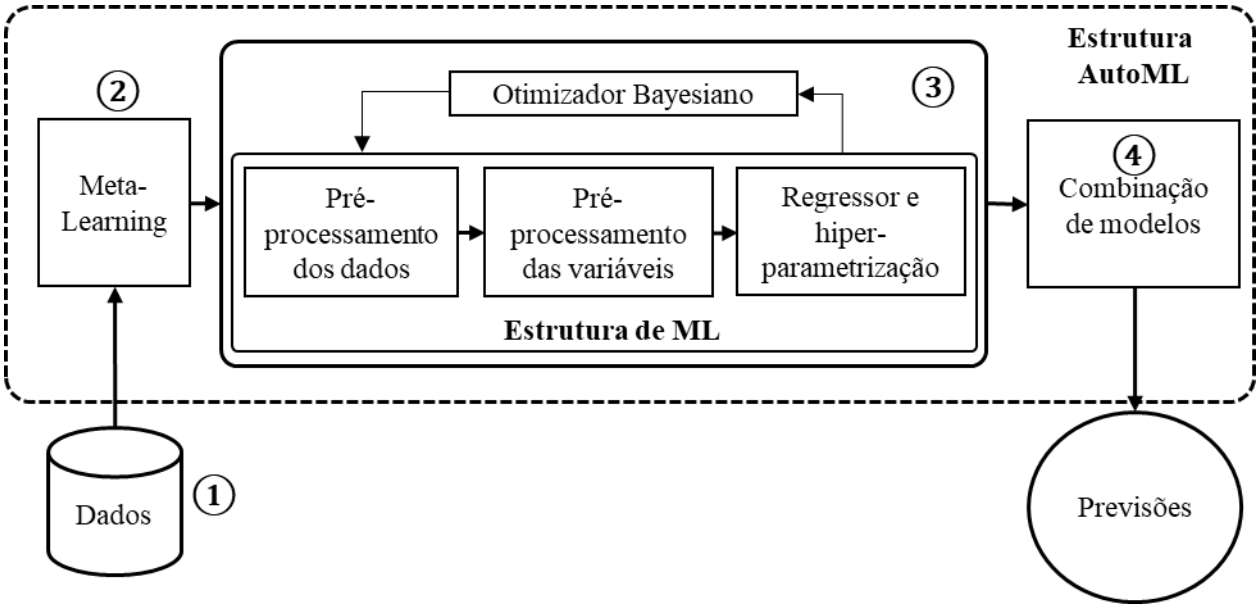
dessa forma o espaço pesquisa de todos os possíveis algoritmos e combinações, tornando a etapa otimização bayesiana mais eficiente;

3) **Otimização Bayesiana.** Este processo inicia com os hiperparâmetros e modelos indicados pela etapa 2 e a partir daí é criado um modelo substituto cujo objetivo é estimar rapidamente o desempenho com um novo conjunto de características candidatas. Então, no ciclo de otimização: **a)** os dados são pré-processados para tratamento, por exemplo, de dados faltantes complementados pela média, mediana ou moda, **b)** depois é selecionado um pré-processador entre os 14 disponíveis para, por exemplo, testar transformações polinomiais das variáveis originais, **c)** é selecionado um algoritmo regressor entre os 12 disponíveis e então o otimizador bayesiano propõe novos hiperparâmetros a partir da seleção inicial até que um limite de performance ou custo seja atingido. Cada conjunto otimizado aqui é chamado de pré-modelo;
4) **Combinação de modelos.** Organizando os pré-modelos a partir daquele com melhor performance, os melhores são combinados em um único modelo de modo a maximizar a performance geral. A performance combinada é geralmente superior em acuraria e redução do *overfitting*, principalmente quando os erros entre os modelos são pouco correlacionados.

A figura 2, adaptada de Feurer *et al*. (2019) apresenta de forma simplificada o fluxo operacional do Auto-Sklearn descrito anteriormente. Importante destacar que a quantidade de pré-processadores, processadores e hiperparâmetros de algoritmos pode atingir um número elevado de fatores a serem testados, daí a importância da estratégia do Auto-Sklearn de iniciar o processo com hiperparâmetros candidatos já selecionados na etapa de meta-learning.

A premissa do algoritmo é a de que um conjunto de dados com características semelhantes terá um desempenho semelhante para um mesmo conjunto de hiperparâmetros. Uma restrição fundamental é o tempo de otimização de cada sub-modelo e do modelo final, e essa restrição é avaliada neste artigo.

Figura 2
Fluxo de Desenvolvimento (Pipeline) do Auto-Sklearn



Adaptado de Feurer *et al.* (2019). A figura apresenta os principais componentes do Auto-Sklearn.

Um dos desafios mais relevantes dos modelos de AutoML é explicar como o modelo funciona, ou seja, identificar os fatores mais relevantes e como eles determinam a variável de interesse, por isso eles são geralmente chamados de modelos *black-box* (Escalante, 2020). Não existe uma solução definitiva, mas entre as ferramentas que apoiam a interpretação desses modelos EBA (2021) destaca: gráficos de dependência parcial, medidas de importância da variável, valor de Shap, Lime (do inglês, *local interpretable model-agnostic explanations*) e explicações contrafactuais, que testam a sensibilidade das estimativas a pequenas variações das variáveis explicativas.

Para comparar os resultados dos modelos tradicionais com o obtido pelo AutoML, optou-se pelo uso do valor de Shap (*Shapley Additive Explanations*) proposta por Lundberg e Lee (2017) e implementada em Python[10]. De forma resumida, trata-se de um método de atribuição aditiva de características, ou seja, ele atribui um peso para cada variável representando a sua importância na estimativa de cada observação, de tal forma que a soma desses pesos se aproxima do valor estimado da variável resposta da respectiva observação em relação ao seu valor médio. Na seção de metodologia são explorados mais detalhes dessa métrica de avaliação.

[10] Para detalhes https://shap-lrjball.readthedocs.io/en/latest/generated/shap.KernelExplainer.html

3. Metodologia

Face à baixa complexidade do produto empréstimo pessoal, nesse estudo nos concentramos nas características da operação e do devedor. O ciclo econômico não é explorado pois a amostra consiste em apenas 4 anos de acompanhamento, e o arcabouço legal não é relevante pois não comparamos jurisdições ou produtos distintos e, além disso, a diferenciação por indústria não se aplica. Finalmente o produto não possui qualquer tipo de garantia real, o que simplifica muito seu cálculo pois toda a recuperação ocorre diretamente em termos financeiros.

Modelos não paramétricos possuem tendência maior ao *overfitting* (Shuermann, 2004), ou seja, uma performance melhor na amostra de desenvolvimento em relação a de teste (Qi e Zhao, 2011), por isso o resultado relevante em termos de performance é aquele obtido nas amostras de teste. O mesmo problema de *overfitting* é observado em modelos que utilizam AutoML (Zöller e Huber, 2021). BIS (2005), Scandizzo (2016) e EBA (2021) destacam que além da amostra de teste também é necessária a avaliação do modelo em uma amostra selecionada fora do período de desenvolvimento, usualmente chamada de *out-of-time*. Essa validação é importante tanto para avaliar o impacto e consistência das variáveis macroeconômicas quando utilizadas quanto para garantir a performance do modelo frente a mudanças estruturais da economia o do setor de exposição do crédito.

Neste artigo a regressão simples e a árvore de decisão são classificadas como metodologia tradicional, conforme explicado anteriormente. Para o desenvolvimento desses modelos foi considerando um subconjunto de variáveis, relevantes teorica ou empiricamente, já apresentadas em estudos anteriores e detalhadas na tabela 2. O objetivo é a identificação clara das variáveis relevantes e a performance do modelo construído sob essas restrições. Para a seleção das variáveis do modelo de regressão linear, apesar da limitação do método (Greenland *et al*, 2016) mas em linha com a metodologia aplicada nesse tipo de análise, foi considerado o p-valor de 6% para seleção das variáveis relevantes e um VIF (*Variance Inflation Factor*) limitado a 11 para mitigar o problema de multicolinearidade.

Para a árvore de decisão foram impostas apenas duas restrições, o número máximo de níveis foi limitado à 10 e a amostra mínima em cada nó folha foi definida em 2,5% da amostra de desenvolvimento, viabilizando assim a interpretação dos resultados. O LGD estimado nessas condições refere-se ao LGD médio de cada nó folha obtido na amostra

de desenvolvimento. Segundo Qi e Zhao (2001) a grande vantagem dessa metodologia em relação a regressão linear é que ela lida melhor com as relações não lineares entre as variáveis, no entanto, árvores muito complexas tendem a apresentar uma performance inferior nas amostras de teste.

Com relação ao AutoML, a parametrização do Auto-Sklearn se limitou ao tempo de processamento do algoritmo, que é objeto de teste neste artigo, e ao número máximo de sub-modelos combinados no modelo final que foi limitado a 5. Quanto a métrica de otimização, também foi utilizada o R^2. Não foi incluído mais nenhum requisito, dando assim a máxima flexibilidade para o algoritmo procurar as soluções ótimas com a mínima intervenção humana possível. Neste cenário uma limitação do algoritmo é a de que os resultados não são reproduzíveis com exatidão pois no processo de otimização os recursos computacionais determinam também os momentos de parada, o que pode gerar modelos diferentes em cada simulação. Dessa forma todas as simulações utilizando o Auto-Sklearn foram realizadas 10 vezes em cada cenário e neste artigo é apresentado e discutido o resultado do modelo de AutoML com o pior e o melhor desempenho.

Todos os modelos foram desenvolvidos usando os dados de empréstimos realizados nos anos de 2014 e 2015, onde 80% foram usados para treinamento dos modelos e 20% foram separados para teste fora da amostra. Além disso, para avaliar a estabilidade e consistência dos modelos, os anos de 2016 e 2017 foram também utilizados como teste, neste caso *out-of-time.*

O desenvolvimento das análises estatísticas foi realizado na linguagem Python utilizando o Google Colaboratory[11]. Os respectivos códigos, bibliotecas e referência a base de dados estão disponíveis no GitHub[12]. As métricas utilizadas para avaliar a performance dos modelos se baseiam naquelas propostas por Scandizzo (2016) e Lotterman (2012) e mensuram tanto a calibração, ou seja, o quão próximos estão dos valores estimados dos observados, quanto a discriminação, ou seja, a ordenação entre os valores observados e estimados, cujo detalhamento é apresentado na tabela 1. Apesar da tabela indicar o pior R^2 com o valor zero, ele pode apresentar limite inferior negativo nas amostras de teste, principalmente nos casos em que há variáveis desnecessárias no modelo

[11] O ambiente é totalmente virtual e está disponível para uso em https://colab.research.google.com/.
[12] https://github.com/Douglasbpinheiro/LGD

e quando o valor médio observado da resposta na amostra de teste muda muito em relação à amostra de desenvolvimento (Bazarbash, 2019).

Tabela 1
Métricas de Performance e suas Características

Métrica	Descrição	Pior	Melhor	Tipo
MAE	*Mean Absolut Error.* Mensura a média da diferença absoluta entre os valores observados e estimados (erros)	$+\infty$	0	Calibração
RMSE	*Root Mean Squared Error*. Refere-se à média da raiz quadrada entre os valores observados e estimados (erros)	$+\infty$	0	Calibração
R^2	*Coeficiente de Determinação* refere-se à proporção da variância que pode ser explicada pelo modelo	0	1	Calibração
r	*Pearson* mede o grau de relação linear entre os valores observados e estimados	0	1	Discriminação
ρ	*Spearman* mede o grau de relação entre os valores observados e estimados usando uma função monotônica.	0	1	Discriminação
T	*Kendall* mede o grau de correspondência entre os valores observados e estimados	0	1	Discriminação

A tabela descreve as métricas utilizadas nesse artigo para avaliar a performance dos modelos. **Métrica**: indicador utilizado na avaliação. **Descrição**: breve explicação sobre o indicador. **Pior e Melhor**: indica os valores críticos do indicador, bem como a forma de interpretação dele. **Tipo**: classifica o indicador em calibração, ou certa, acerto da estimativa, e discriminação, ou seja, associação ou ordenação da estimativa em relação aos valores observados.

Na metodologia tradicional como proposta aqui a identificação dos fatores e seus efeitos é clara, assim para comparar as variáveis selecionadas por essa metodologia com aquelas de maior relevância do modelo de AutoML foi utilizado o método Kernel Shap. Nesse método de avaliação os pesos das variáveis são obtidos a partir de uma regressão linear ponderada, onde a variável testada é retirada da equação e o efeito da sua ausência é observado na resposta, sendo o processo repetido com todas as variáveis utilizadas na amostra de treinamento. Apesar de ser mais eficiente que outras métricas, esse método ainda é computacionalmente oneroso (Lundberg e Lee, 2017), dessa forma as amostras para cálculo dos valores de Shap de todas as simulações foram limitadas a 500 observações.

Importante destacar que o valor de Shap se trata de uma medida local para a respectiva observação e não significa necessariamente que o peso causa o valor estimado, mas apenas que para uma determinada observação, aquele peso contribuiu para compor a estimativa da variável de interesse. Desta forma, quando o valor de Shap é próximo de zero, significa que a variável contribui muito pouco para a estimativa. Para uma interpretação global a forma mais comum é a utilização da somatória do módulo do valor de Shap das observações avaliadas, que é utilizado aqui para identificar as variáveis mais relevantes nos diferentes modelos (Gianfagna e Di Cecco, 2021). A amostra foi selecionada aleatoriamente em todas as simulações, sendo utilizada a opção de processamento que considera a existência de correlação entre as variáveis explicativas, o que de fato se mostra muito relevante quando considerado o alto VIF observado nos resultados da regressão linear.

Com relação às variáveis relevantes para a modelagem do LGD, apesar da discussão a respeito da identificação na seção de resultados e sua relação com a literatura existente, nenhuma hipótese sobre esse tema é proposta. Espera-se apenas que se a métrica de análise das variáveis relevantes testada for consistente, será observada uma similaridade entre as variáveis escolhidas pelos métodos tradicionais e automáticos.

Dessa forma, as hipóteses testadas se concentram na questão da performance comparativa dos modelos, conforme segue:

Hipótese 1: Modelo de Auto Machine Learning têm performance superior aos Modelos Tradicionais para a estimativa de LGD.

A primeira hipótese parte da vasta literatura apresentada aqui, onde foi constatado que modelos de ML apresentam performance superior, principalmente por capturarem efeitos não observados ou desconhecidos da estrutura de dados de forma mais eficiente que os modelos paramétricos. Assim, é utilizado nessa avaliação o modelo de AutoML, onde toda a hiper-parametrização, escolha do modelo e sua combinação ocorre de forma automática, ou seja, sem nenhum pressuposto a respeito da estrutura de determinantes do LGD.

Hipótese 2: A performance do Modelo de Auto Machine Learning é melhor quanto maior a amostra de treino.

Quanto à segunda hipótese, o uso de algoritmos complexos para a identificação de padrões é mais preciso quanto maior o número e a variabilidade de informações analisadas (Lundberg e Lee, 2017; EBA, 2021), daí a intensa utilização do ML na era do *Big-Data*. Da mesma forma, no AutoML, que possui maior dependência do aprendizado do algoritmo, a expectativa é a de que o modelo tenha uma performance melhor quanto maior o número de observações utilizada no desenvolvimento. Para a modelagem de LGD isso é especialmente importante, pois muitas bases de dados são limitadas em número de observações, principalmente em carteiras com baixo nível de inadimplência, por exemplo, no segmento de grandes empresas, o que pode inviabilizar o uso da metodologia de AutoML em segmentos específicos. Por exemplo Qi e Zhao (2011) trabalham com uma amostra de 3751 observações considerando o período de 1985 a 2008, o que representa cerca de 163 observações por ano e da mesma forma Yashkir e Yashkir (2013) utilizam dados de 4275 observações da S&P de 1981 a 2011, o que representa apenas 142 observações por ano, porém muitas vezes as empresas não dispõem de um volume ou histórico abrangente como esses, então avaliar a eficácia da metodologia em amostras menores torna-se ainda mais relevante. A única maneira de aumentar o volume desse tipo de dado é aguardar e coletar novos dados ao longo do tempo (Israel *et al*., 2020).

Para testar essa hipótese, da base original de treinamento foram realizadas as seguintes simulações: a) toda a amostra disponível para treinamento foi utilizada, ou seja, 97.147 observações, b) 10% da amostra foi treinada, perfazendo 9.714 observações e, c) 0,5% da amostra foi treinada, representada por 485 observações. A partir desses conjuntos poderemos avaliar a performance comparativa do modelo avançado (AutoML) e tradicional considerando a questão da limitação amostral.

Hipótese 3: A performance do Modelo de Auto Machine Learning é melhor quanto maior o tempo de treino.

A técnica de AutoML é onerosa em termos de consumo de recursos computacionais, portanto a avaliação do período de otimização pode ser um diferencial quando há a necessidade de treinamento de vários modelos simultaneamente. Para avaliar a hipótese de que um maior tempo de treinamento melhora a performance do algoritmo os modelos foram treinados considerando 2 restrições de tempo: a) 3 horas de tempo total e 20 minutos por sub-modelo e, b) 1 hora de tempo total e 7 minutos por sub-modelo.

Nos modelos desenvolvidos na metodologia tradicional essa restrição não é relevante pois os algoritmos e hiperparâmetros foram previamente definidos.

4. Dados e Variáveis

Esta seção descreve a fonte de informação utilizada e apresenta uma discussão sumária dos dados e tratamentos.

4.1. Dados e seleção da amostra

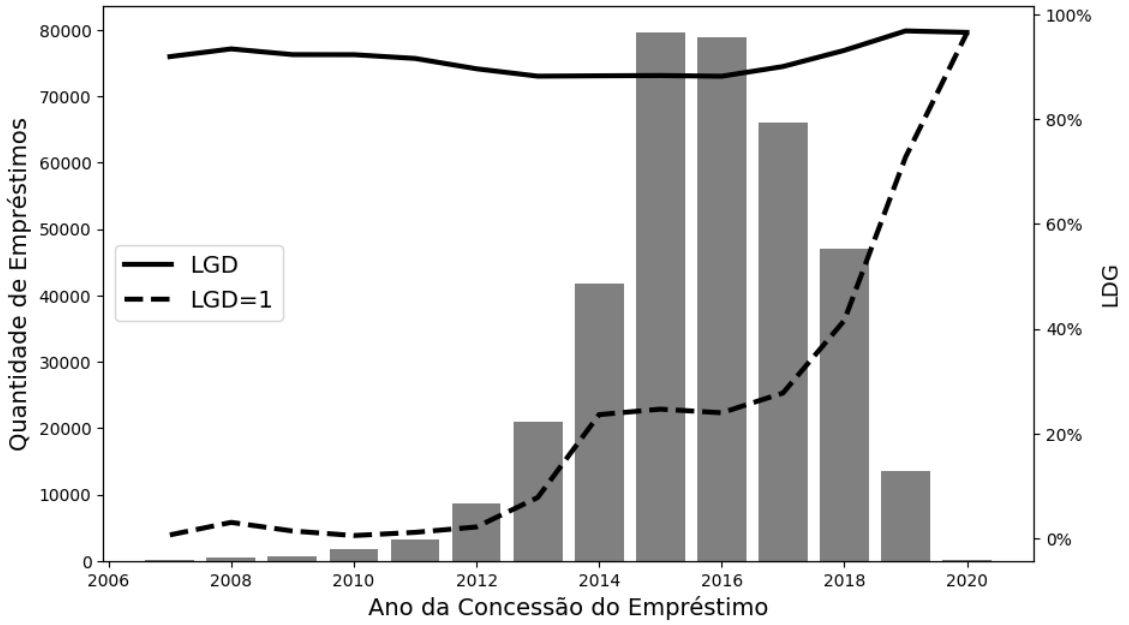
A base com as informações foi obtida do Kaggle[13] e possui 141 variáveis e 2.925.493 empréstimos realizadas entre 2007 e 2020 pelo *Lending Club*. Utilizando o mesmo critério de Zhou *et al.* (2018) foram selecionados todos os empréstimos na condição de baixados por inadimplência, o que normalmente ocorre 120 dias após o vencimento da parcela do empréstimo. No entanto, diferente de Zhou *et al.* (2018) e para melhorar a apuração do LGD foi incluído o custo de recuperação cobrado pelo *Lending Club* dos investidores, que representa em média 17,5% da recuperação observada. O LGD, portanto, foi definido considerando o seguinte cálculo:

$$LGD = 1 - \left[\frac{(Valor\ Recuperado - Custo\ de\ Cobrança)}{(Valor\ do\ Empréstimo - Principal\ Pago)} \right] \qquad (2)$$

A figura 3 apresenta a quantidade de empréstimos baixados por inadimplência anualmente entre 2006 e 2020, o que soma 363.309 operações. O LGD médio gira em torno de 90%, porém a quantidade de registros com LGD igual a 100% (ou seja, aqueles casos em que não se observa nenhuma recuperação) aumenta significativamente após 2018 em função do menor período de apuração da recuperação, e é significativamente menor no período anterior a 2014, quando a quantidade de empréstimos era ainda pequena, portanto, os empréstimos realizados antes de 2014 e após 2017 foram excluídos da análise pois não são comparáveis. Também foram excluídos 31 empréstimos cujo valor principal foi integralmente pago, impedindo o cálculo do LGD como proposto aqui, o que resultou em uma amostra final de 266.515 observações no período entre 2014 e 2017, o que representa 73,3% dos empréstimos inadimplentes. Apenas 17 observações atípicas apresentaram LGD < -0,3, mas foram mantidas censurando o valor nesse nível.

[13] O endereço https://www.kaggle.com/code/ztrimus/loan-repayment-prediction/data?select=Loan_status_2007-2020Q3.gzip dá acesso à base de dados utilizada nesse artigo, bem como ao dicionário de variáveis.

Figura 3
Quantidade de Empréstimos e LGD Médio por Ano



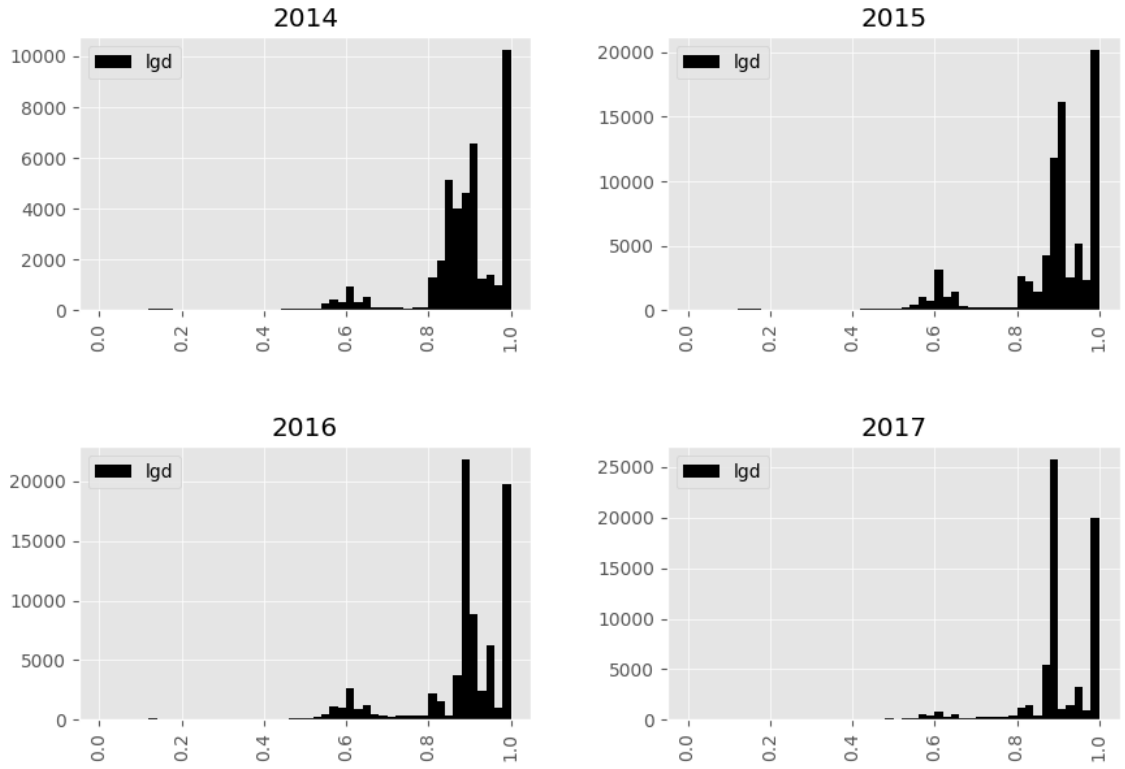
A variável LGD refere-se àquela apresentada na equação 1, já a variável LGD=1 refere-se à proporção de empréstimos que não apresentaram recuperação, ou seja, quando o LGD foi igual a 100%.

Zhou *et al.* (2018) excluem operações cujos clientes não tiveram renda verificada e operações com informações faltantes, reduzindo a amostra em 87%. No caso da renda verificada não há diferença relevante no LGD observado e para os casos de informações faltantes, além de serem pouco representativos também consideramos essas situações no desenvolvimento do estudo, portando todas foram mantidas.

Quanto ao custo do dinheiro no tempo, ele não foi considerado pois não há informação de data de inadimplência e data de recuperação, além disso, o efeito não deve ser relevante pois o valor do LGD se estabiliza em média 2 anos após a data da concessão, conforme observado no gráfico 1, e as taxas de juros no mercado e períodos analisados foram baixas nos Estados Unidos.

A figura 4 mostra o LGD distribuído nos diferentes anos da amostra. Observa-se uma concentração da distribuição à direita, o que é esperado para um produto de varejo sem garantias associadas, o que acarreta um elevado nível de perda após a ocorrência da inadimplência. Além disso, não há variação relevante entre os diferentes anos.

Figura 4
Distribuição do LGD



Distribuição do LGD considerando os anos de concessão do crédito selecionados para o desenvolvimento e teste dos modelos.

4.2. Variáveis

Variáveis com baixa representatividade ou com impedimento do uso por aspectos conceituais foram excluídas. Das 52 exclusões 31 decorreram de baixo preenchimento (preenchimento de até 2,7%), 6 têm o mesmo preenchimento para todos os registros, 7 referem-se a dados futuros e/ou compõem o cálculo do LGD e as últimas 8 apresentaram deficiências diversas. Após a criação de algumas variáveis combinadas e transformação em categorias binárias de algumas variáveis auxiliares a base de dados final manteve 112 variáveis para a aplicação do modelo de AutoML, incluindo o LGD.

No processo de modelagem tradicional, inicialmente são selecionadas informações que possuem sentido econômico sobre o evento analisado. Optou-se aqui por utilizar o mesmo conjunto de variáveis de Zhou *et al.* (2018) apresentadas na tabela 2, mostrando inclusive o efeito esperado sobre o LGD. Além delas, foram incluídas informações sobre o comportamento de crédito do cliente, saldo do empréstimo no momento da inadimplência e o uso declarado dos recursos, todas detalhadas na tabela 2.

Tabela 2
Informações utilizadas na Modelagem Tradicional

Variável	Definição	Sinal
loan_amnt (1)	Valor do empréstimo.	(+)
term (1)	Assume 1 quando o prazo do empréstimo é de 60 meses e 0 quando o prazo é 36 meses.	(+)
int_rate (1)	Taxa anual de juros do empréstimo.	(+)
grade (1)	Classificação de crédito definida na concessão, separado em 7 classes de A (menor risco de inadimplência) a G (maior risco) e referência.	(-)
ead	Valor do saldo do empréstimo no momento da inadimplência.	(+)
ead_inc	Valor do saldo do empréstimo no momento da inadimplência dividido pela renda anual do devedor.	(+)
ead_loan	Valor do saldo do empréstimo no momento da inadimplência dividido pelo valor contratado.	(+)
emp_length (1)	Tempo no emprego do devedor medido em anos, limitado a 0,5 se for menor que 1 ano e 10 se for maior.	(-)
home_ownership (1)	Classifica a moradia em (0) Alugada "rent" adotada como referência, (1) Financiada *"mortgage"*, (2) Outros *"others"* e (3) Própria *"own"*.	(-)
annual_inc (1)	Renda anual do devedor.	(-)
dti (1)	Comprometimento mensal de renda (parcela/salário) excluindo o empréstimo que está sendo analisado e eventual crédito imobiliário contratado.	(+)
delinq_2yrs (1)	Número de atrasos acima de 30 dias nos últimos 2 anos	(+)
mths_since_last_delinq	Número de meses desde que ocorreu o último atraso > 30 dias pelo devedor.	(-)
mths_since_last_delinq_null	Indicador assume 1 quando não há histórico de atraso do devedor.	(-)
pub_rec	Número de registros negativos do devedor no mercado.	(+)
inq_last_12m	Número de consultas de crédito do devedor nos últimos 12 meses.	(+)
purpose	propósito do empréstimo: carro, cartão de crédito, consolidação de dívidas *"debt consolidation"*, educação, reforma *"home improvement"*, casa *"house"*, compra de alto valor, médico, mudança *"moving"*, energia renovável, férias, casamento, outros *"other"* e pequenos negócios *"small business"*, adotada como referência.	Não se aplica
fico_range_high (low)	Limite superior (inferior) do escore Fico(2) do devedor.	(-)

Variável indica o nome da variável utilizada na avaliação tradicional. **Definição** traz uma breve explicação sobre a construção da variável. **Sinal** indica se o efeito esperado da variável sobre o LGD é positivo (aumenta-piora) ou negativo (diminui-melhora). **(1)** Variáveis também utilizadas por Zhou *et al*. (2018). **(2)** https://www.fico.com/en/products/fico-score

Na medida do possível, os nomes das variáveis foram mantidos da forma que estão disponibilizados na base de dados original, facilitando a eventual identificação e validação dos resultados. Para a regressão linear, as variáveis *loan_amnt* e *annual_inc* foram divididas por 1000 e como são relativamente assimétricas foi aplicado o logaritmo natural nos seus valores para as regressões, o mesmo feito com a variável *int_rate*. Outras transformações são possíveis, mas isso estava fora do escopo desse artigo. Na árvore de decisão elas foram utilizadas sem qualquer tratamento.

Além disso, foram criadas as variáveis *ead*, *ead_inc* e *ead_loan* que não fazem parte do conjunto original de informações, no entanto elas são muito importantes pois representam o saldo devedor do cliente em termos absolutos, em função da renda e em função do valor originalmente emprestado e são apuradas no momento da inadimplência. Importante destacar que quanto maior o *ead* e o *ead* em função da renda mais difícil se torna a recuperação do crédito face ao maior volume e representatividade da dívida. No caso do *ead_loan*, um maior valor significa que a inadimplência ocorreu próximo do momento da concessão do empréstimo, o que também indica uma dificuldade maior do devedor e, portanto, uma menor expectativa de recuperação. As variáveis criadas não foram incluídas no modelo de AutoML, pois essas possíveis combinações são parte intrínseca do algoritmo, ou seja, se o algoritmo for eficaz e as variáveis forem relevantes, elas serão identificadas, combinadas e utilizadas no modelo de forma automática.

Todas as demais informações derivam da base de dados original e alguns tratamentos complementares foram: 1) censura na variável ead_inc no valor 30 que impactou apenas 89 observações, 2) censura na variável dti em 200 que impactou 56 observações, 3) censura na variável annual_inc para valores menores de 100 e maiores de 1.000.000 que impactou 87 observações, 4) os campos nulos foram substituídos por zero e 5) foram criadas 5 variáveis complementares que indicavam essas substituições, um exemplo é a *mths_since_last_delinq* onde um maior valor significa que o devedor apresentou um atraso há muito tempo, o que é um bom indicativo em relação àqueles que inadimpliram recentemente, enquanto o 0 (zero) significa que não há atraso no histórico, o que também é um bom indicativo. Dessa forma o efeito é não linear e a variável binária complementar *mths_since_last_delinq_null* busca capturar esse efeito. Apesar de usados em todas as simulações para comparabilidade, esses últimos tratamentos são importantes

principalmente para a análise de regressão linear, mas desnecessários para os demais algoritmos.

Os sinais esperados da tabela 2 são discutidos na seção de resultados apenas no caso daquelas que foram relevantes. Finalmente, apesar de não ser requisito técnico para o AutoML, os tratamentos apresentados, salvo exceções destacadas, foram os mesmo para todos os modelos. Isso porque o desenvolvimento de modelos no mercado financeiro pressupõe um conhecimento adequado e justificado das informações utilizadas, assim os resultados obtidos aqui são comparáveis com o ambiente de uso, além disso, isso torna a comparação mais adequada entre os modelos ao uniformizar os dados de entrada. Na próxima seção são apresentados os resultados dos testes realizados.

5. Resultados

Nesta seção são discutidos os resultados das três hipóteses apresentadas e depois são comparadas as variáveis mais relevantes em cada modelo. A tabela 3 apresenta os resultados do uso da metodologia tradicional e as tabelas 4 e 5 da metodologia avançada.

Para a avaliação da primeira hipótese, ou seja, a respeito da superioridade da metodologia avançada na performance, nos fixamos na performance das amostras de teste que utilizaram 100% dos dados disponíveis para o desenvolvimento. Na metodologia tradicional os resultados da tabela 3 mostram que não há diferença relevante entre a performance entre árvore de decisão (painel A) e regressão linear (painel B). No entanto, quando comparados os resultados com a tabela 4, onde foi utilizado o AutoML com apenas 1 hora de processamento (cenário mais restrito do que aquele com 3 horas de duração), a diferença é muito relevante.

Os valores do MAE e RMSE da amostra de teste de 2 anos à frente da árvore de decisão (regressão linear) caem de 0,07 (0,069) e 0,11 (0,11) para 0,015 e 0,029 respectivamente na tabela 4 (Painel A – Valor Máximo), ou seja, os erros são reduzidos em pelo menos de 70% na pior simulação da metodologia avançada. Quanto às demais métricas o impacto é até mais relevante, pois elas saem de valores negativos da tabela 3 para valores muito significativos na tabela 4.

Em geral os indicadores de erro em todas as amostras de teste são pelo menos 50% menores no método de AutoML com apenas 1 hora de processamento comparado ao método tradicional. A menor diferença entre os resultados foi no RMSE da amostra de

teste de 0,5% da população, cujo indicador apresentou erro 36% (33%) menor que o obtido na Árvore de Decisão (Regressão Linear). Considerando as métricas de ordenação, a diferença é ainda significativa, em muitos casos melhorando cerca de 10 vezes o indicador na metodologia avançada de AutoML.

Tabela 3 (A)
Método Tradicional – Árvore de Decisão

Tamanho	Amostra	MAE	RMSE	R^2	r	ρ	τ
100%	Desenvolvimento	0,0872	0,1301	0,0119	0,109	0,0676	0,0471
	Teste	0,0874	0,1313	0,0105	0,1026	0,0656	0,0454
	Teste 1 ano	0,0845	0,1296	0,0015	0,0625	0,0854	0,0599
	Teste 2 anos	0,0703	0,1096	-0,05	-0,059	-0,062	-0,044
10%	Desenvolvimento	0,0881	0,1306	0,0295	0,1719	0,1392	0,097
	Teste	0,0927	0,1397	-0,009	0,0562	0,0349	0,0244
	Teste 1 ano	0,0867	0,1308	-0,016	0,0379	0,072	0,0499
	Teste 2 anos	0,0726	0,1108	-0,072	-0,027	0,0048	0,0036
0,5%	Desenvolvimento	0,0791	0,1141	0,2	0,4473	0,4023	0,2927
	Teste	0,1021	0,1392	-0,188	0,0481	-0,025	-0,012
	Teste 1 ano	0,0998	0,1417	-0,193	0,0118	0,026	0,019
	Teste 2 anos	0,0882	0,1238	-0,339	0,0125	0,022	0,0168

Tabela 3 (B)
Método Tradicional – Regressão Linear

Tamanho	Amostra	MAE	RMSE	R^2	r	ρ	τ
100%	Desenvolvimento	0,0871	0,1304	0,008	0,0898	0,04	0,0278
	Teste	0,0871	0,1314	0,008	0,0883	0,0496	0,0344
	Teste 1 ano	0,0839	0,1294	0,005	0,0727	0,0819	0,057
	Teste 2 anos	0,0689	0,1088	-0,035	-0,032	-0,049	-0,033
10%	Desenvolvimento	0,0882	0,1317	0,0095	0,0977	0,0806	0,055
	Teste	0,0899	0,1399	-0,001	0,0438	0,0431	0,0292
	Teste 1 ano	0,0862	0,1384	-0,139	0,0173	0,0757	0,0514
	Teste 2 anos	0,0753	0,1662	-1,412	0,0078	0,0124	0,0085
0,5%	Desenvolvimento	0,0817	0,1246	0,0239	0,1547	0,162	0,0135
	Teste	0,0875	0,1344	-0,019	-0,043	-0,064	-0,054
	Teste 1 ano	0,0851	0,131	-0,02	0,004	0,0089	0,0074
	Teste 2 anos	0,0688	0,11	-0,057	0,001	0,0026	0,0022

A tabela apresenta os resultados do modelo desenvolvido com diferentes tamanhos amostrais e testado em amostra de 20% extraída da base de desenvolvimento, ou *out-of-sample* (teste), e em amostra *out-of-time* de dados 1 ano à frente (teste 1 ano) e outra 2 anos à frente (teste 2 anos). Os indicadores são detalhados na tabela 1. O painel A apresenta o resultado da performance da Árvore de Decisão e o painel B a performance da Regressão Linear desenvolvidos utilizando a metodologia tradicional.

O resultado robusto na menor amostra, constituída de apenas 485 observações de treino, reforça a importância do uso de uma ampla gama de variáveis e metodologias

mesmo em um cenário limitado de dados. Esse resultado torna evidente a relação não linear entre as variáveis que não é capturada pelos métodos tradicionais (Israel *et al.*, 2020). O conjunto desses resultados corrobora dessa forma clara a primeira hipótese, não restando dúvida em relação à superioridade do AutoML em relação às metodologias tradicionais.

Independente da metodologia escolhida, ao não explorar todo o conjunto de variáveis disponíveis muita informação sobre as relações é perdida, o que torna evidente a limitação da metodologia tradicional ao reduzir o conjunto inicial de aprendizado às variáveis com sentido econômico já propostos pela literatura, limitando assim o aprendizado dos algoritmos, sejam eles complexos ou não. Além disso, como veremos a seguir, a identificação de variáveis relevantes pós treino pode trazer novos insights e linhas de investigação sobre o fenômeno estudado, desconhecidos até então.

Quanto à segunda hipótese, ela propõe que a performance do modelo de AutoML cai à medida que a amostra de treino é reduzida. A tabela 4 mostra que em 1 hora de processamento quando a amostra é reduzida para 10% da original a média do erro absoluto no critério de valor máximo aumenta 76% e 88% respectivamente nas amostras de teste 1 ano e 2 anos à frente respectivamente, enquanto o erro quadrado aumenta 47% e 63% respectivamente. Quando a amostra é reduzida de 10% para 0,5% esses erros no critério de valor máximo aumentam cerca de 40% nos mesmos cenários. Se considerados também os valores mínimos os resultados são semelhantes. Quanto ao R^2 nas amostras de teste 1 e 2 anos à frente, esse piora cerca de 10% quando a amostra é reduzida para 10% e piora cerca de 25% quando a amostra é reduzida de 10% para 0,5%.

Finalmente a maioria das métricas de ordenação apresentam piora nas amostras de teste *out-of-time* quando ocorreu a redução do conjunto de teste do processamento em 1 hora. A única exceção relevante foi na redução de 10% para 0,5% da amostra no critério valor máximo que apresentou melhora da correlação de *Spearman* e *Kendall* de 4,5% e 7,1% respectivamente, o que indica que para esse conjunto de dados o aumento do volume de observações não incremente necessariamente todos os indicadores de ajuste modelo.

Já na tabela 5, onde o tempo de processamento é maior, quando a amostra foi reduzida de 100% para 10% os indicadores de erro melhoraram sensivelmente, chegando a uma redução de até 47% de nas amostras de testes, bem como o R^2 que aumentou até 12,3% e na ordenação que melhorou até 11%. Inicialmente esse resultado parece contraintuitivo, porém vemos uma estabilidade ou piora nesses mesmos indicadores

quando comparamos as amostras de desenvolvimento, indicando um efeito de *overfitting* nas simulações com 3 horas de duração em toda a amostra de desenvolvimento. Agora quando a amostra foi reduzida de 10% para 0,5%, os resultados corroboram a hipótese testada aqui de forma inequívoca. Nas amostras de teste os erros MAE e RMSE aumentaram entre 113,7% e 308%, o R^2 variou de -33,5% a -46,1% e os indicadores de ordenação variaram entre -6,9% e -24,6%. Dessa forma a segunda hipótese é corroborada por todas as métricas, reforçando a importância de desenvolver o modelo com a maior amostra de treino possível no contexto de métodos avançados de análise. Cuidado deve ser tomado com o tempo de treino, para mitigar o risco de overfitting.

Finalmente, quanto à última hipótese, a análise deve ser feita comparando as tabelas 4 e 5, com 1 e 3 horas de processamento respectivamente. Para a amostra de treino completa as performances medidas pelo MAE e RMSE de fato melhoraram entre 22% e 50% no critério de valor máximo nas amostras *out-of-time* quando o tempo de treino foi limitado a 1 hora, contrariado a hipótese avaliada. No entanto o resultado piorou na amostra de teste *out-of-sample* e nas simulações pelo critério valor mínimo, conforme esperado. Os resultados na amostra de 10% foram aqueles esperados, com o erro aumentando significativamente na amostra com 1 hora de processamento e o R2 e indicadores de ordenação piorando em todos os cenários e métricas. Finalmente na menor amostra não foram observadas diferenças relevantes em nenhuma métrica analisada.

De forma geral os resultados foram mistos e não foi possível corroborar a terceira hipótese, ou seja, não observamos diferença significativa de performance em todos os cenários quando o treinamento foi reduzido de 3 horas para 1 hora. Esses resultados são similares aos observados por Zoller e Huber (2021), que testando vários algoritmos de AutoML não observaram um ganho relevante de performance quando alteraram o tempo de processamento de 1 hora para 4 ou 8 horas de processamento.

Em aplicações práticas no mercado financeiro um tempo de treino reduzido é importante para que múltiplos cenários possam ser testados, portanto o resultado observado neste estudo fornece aos usuários de AutoML referência e segurança relevante para a sua utilização em ciclos mais curtos de processamento do modelo.

Tabela 4

Painel A - AutoML – Processamento 1 hora – Amostra 100%

Tamanho	Amostra	MAE	RMSE	R^2	r	ρ	τ
Valor Máximo	Desenvolvimento	0,0133	0,028	0,967	0,9841	0,9682	0,8761
	Teste	0,0186	0,0394	0,9207	0,9607	0,9416	0,8301
	Teste 1 ano	0,0184	0,0391	0,9179	0,9589	0,9225	0,7955
	Teste 2 anos	0,0152	0,0291	0,9349	0,9681	0,8857	0,7267
Valor Mínimo	Desenvolvimento	0,0104	0,0238	0,9543	0,978	0,9609	0,8609
	Teste	0,017	0,0372	0,9111	0,9567	0,9313	0,8144
	Teste 1 ano	0,017	0,0372	0,9091	0,9548	0,9102	0,7782
	Teste 2 anos	0,014	0,0273	0,9259	0,9642	0,8701	0,7064

Painel B - AutoML – Processamento 1 hora – Amostra 10%

Tamanho	Amostra	MAE	RMSE	R^2	r	ρ	τ
Valor Máximo	Desenvolvimento	0,024	0,0469	0,9377	0,9691	0,953	0,8447
	Teste	0,0336	0,0671	0,8375	0,9199	0,9059	0,7631
	Teste 1 ano	0,0325	0,0577	0,8338	0,915	0,8601	0,7021
	Teste 2 anos	0,0286	0,0475	0,8374	0,9174	0,8223	0,6451
Valor Mínimo	Desenvolvimento	0,015	0,0331	0,875	0,9373	0,9142	0,7813
	Teste	0,0282	0,0561	0,7671	0,881	0,866	0,7161
	Teste 1 ano	0,0298	0,0528	0,8021	0,898	0,842	0,6864
	Teste 2 anos	0,0257	0,0431	0,8019	0,8988	0,7953	0,6176

Painel C - AutoML – Processamento 1 hora – Amostra 0,5%

Tamanho	Amostra	MAE	RMSE	R^2	r	ρ	T
Valor Máximo	Desenvolvimento	0,0426	0,0705	0,7033	0,8393	0,8705	0,7189
	Teste	0,0482	0,0896	0,5129	0,7351	0,8777	0,7197
	Teste 1 ano	0,047	0,0811	0,6193	0,7899	0,8696	0,7073
	Teste 2 anos	0,0397	0,0677	0,612	0,7828	0,8595	0,6911
Valor Mínimo	Desenvolvimento	0,0417	0,0695	0,6944	0,8339	0,8563	0,6955
	Teste	0,0474	0,0892	0,5087	0,7306	0,8464	0,68
	Teste 1 ano	0,0461	0,08	0,6092	0,7834	0,843	0,6673
	Teste 2 anos	0,0384	0,0667	0,5992	0,7751	0,7897	0,6048

A tabela apresenta os resultados do modelo desenvolvido com 1 hora de processamento e com diferentes tamanhos amostrais, testado em amostra de 20% extraída da base de desenvolvimento, ou *out-of-sample* (teste), e em amostra *out-of-time* de dados 1 ano à frente (teste 1 ano) e outra 2 anos à frente (teste 2 anos). Os indicadores são detalhados na tabela 1. Foram realizadas 10 simulações em cada cenário e o Valor Máximo (Mínimo) refere-se ao maior (menor) valor observado do respectivo indicador.

Tabela 5

Painel A - AutoML – Processamento 3 horas – Amostra 100%

Tamanho	Amostra	MAE	RMSE	R^2	*r*	ρ	τ
Valor Máximo	Desenvolvimento	0,0129	0,0276	0,9743	0,9873	0,9711	0,8858
	Teste	0,0178	0,0375	0,9447	0,973	0,9528	0,8478
	Teste 1 ano	0,0316	0,0504	0,933	0,9662	0,9264	0,8011
	Teste 2 anos	0,0304	0,0454	0,9471	0,9739	0,8849	0,7273
Valor Mínimo	Desenvolvimento	0,0085	0,021	0,9557	0,9782	0,9598	0,8582
	Teste	0,0145	0,0311	0,9192	0,9603	0,9332	0,8188
	Teste 1 ano	0,0154	0,0336	0,8489	0,9233	0,8668	0,7066
	Teste 2 anos	0,0127	0,0246	0,8201	0,9083	0,8152	0,6377

Painel B - AutoML – Processamento 3 horas – Amostra 10%

Tamanho	Amostra	MAE	RMSE	R^2	*r*	ρ	τ
Valor Máximo	Desenvolvimento	0,0128	0,0301	0,9577	0,9792	0,971	0,8789
	Teste	0,0136	0,0344	0,95	0,9752	0,9695	0,8791
	Teste 1 ano	0,0188	0,038	0,9428	0,9716	0,9526	0,8472
	Teste 2 anos	0,0162	0,0301	0,9543	0,9781	0,9236	0,7856
Valor Mínimo	Desenvolvimento	0,0109	0,0273	0,9486	0,9743	0,9571	0,8527
	Teste	0,0116	0,0311	0,9389	0,9694	0,9588	0,8557
	Teste 1 ano	0,0117	0,031	0,914	0,9575	0,915	0,7813
	Teste 2 anos	0,0099	0,0229	0,9211	0,9604	0,871	0,7078

Painel C - AutoML – Processamento 3 horas – Amostra 0,5%

Tamanho	Amostra	MAE	RMSE	R^2	*r*	*P*	*T*
Valor Máximo	Desenvolvimento	0,0421	0,0702	0,7038	0,8396	0,8709	0,7194
	Teste	0,0484	0,0895	0,5122	0,7354	0,8785	0,7211
	Teste 1 ano	0,0467	0,0812	0,618	0,7888	0,8697	0,7075
	Teste 2 anos	0,0395	0,0679	0,6103	0,7818	0,86	0,6918
Valor Mínimo	Desenvolvimento	0,0417	0,0694	0,6975	0,8358	0,8576	0,6984
	Teste	0,0474	0,0892	0,5091	0,7306	0,8305	0,6813
	Teste 1 ano	0,0461	0,0801	0,608	0,7826	0,8481	0,6752
	Teste 2 anos	0,0384	0,0668	0,5969	0,7738	0,796	0,6091

A tabela apresenta os resultados do modelo desenvolvido com 3 horas de processamento e com diferentes tamanhos amostrais, testado em amostra de 20% extraída da base de desenvolvimento, ou *out-of-sample* (teste), e em amostra *out-of-time* de dados 1 ano à frente (teste 1 ano) e outra 2 anos à frente (teste 2 anos). Os indicadores são detalhados na tabela 1. Foram realizadas 10 simulações em cada cenário e o Valor Máximo (Mínimo) refere-se ao maior (menor) valor observado do respectivo indicador.

Com relação à performance geral, Zhou *et al*. (2008) utilizando uma amostra semelhante da Lending Club obtiveram um R^2 máximo de 0,0584 na amostra de treinamento, bem superior ao modelo de regressão desse estudo que não aplicou os mesmos filtros, no entanto muito inferior ao modelo de AutoML que na amostra comparativamente próxima de 10% da população atingiu um R^2 de 0,875 no valor mínimo da amostra de treino e 0,802 no valor mínimo da amostra de teste de 2 anos à frente. Papoušková e Hajek (2019) utilizando a técnica de ML de Random Forest em dois

estágios obtiveram um RMSE de 0,053 e um MAE de 0,026. Os resultados comparáveis obtidos no modelo de AutoML com 3 horas de processamento em 100% da amostra de treino (volume de operações comparável entre os 2 estudos) foram RMSE Valor Máximo de 0,0276 e MAE Valor Máximo de 0,0129, ou seja, o AutoML proposto aqui atingiu resultado significativamente superior ao modelo de RF em 2 estágios, com diferença no MAE e RMSE de cerca de 50%. O R^2 também foi superior no pior cenário do AutoML em 2,3%. Importante lembrar que todos os resultados são comparados na amostra de desenvolvimento pois Papoušková e Hajek (2019) não trabalharam com amostra de teste.

A última análise refere-se à seleção das informações relevantes para a estimação do LGD. O principal objetivo é verificar até que ponto a performance superior do modelo de AutoML foi causado pela inclusão de variáveis sem o embasamento econômico prévio ou decorreu da obtenção de melhores combinações entre as variáveis já consideradas relevantes na metodologia tradicional. A tabela 6 apresenta as variáveis mais relevantes observadas na árvore de decisão e cabe destacar a relevância do *ead_loan* e *annual_inc*, ou seja, o valor inadimplido em relação ao empréstimo e a renda anual do devedor, em quase todos os nós iniciais de todos os cenários. As demais apresentaram menor frequência. Cabe lembrar ainda que as variáveis *ead*, *ead_inc* e *ead_loan* não foram utilizadas no modelo de AutoML então não serão diretamente comparáveis entre as metodologias.

A tabela 7 apresenta as variáveis mais relevantes observadas na regressão linear e nela novamente o *ead_loan* apresentou relevância estatística e o sinal esperado, ou seja, quanto maior o valor inadimplido em relação ao empréstimo, maior o LGD. Já a variável *annual_inc* não foi relevante em nenhuma das análises de regressão, ou seja, mesmo entre as metodologias simplificadas não é possível garantir a consistência em relação às variáveis escolhidas. A variável *loan_amnt* relativa ao valor do empréstimo e *fico_range_low*, ou seja, o menor escore do devedor em cada empréstimo, foram relevantes apenas na árvore de decisão.

Quanto às demais, observamos que o *emp_length* apresentou o resultado esperado, ou seja, quanto maior o tempo no emprego do devedor maior a capacidade de regularizar a dívida inadimplida. A variável *term* assume 1 para empréstimos mais longos e, portanto, mais arriscados apresentou o sinal positivo esperado, o mesmo ocorrendo com o *ead_inc* que indica o volume da dívida em relação à renda do devedor. *Pub_rec* indica o número

de registros negativos do devedor no mercado e apresentou o sinal positivo esperado. Quanto à moradia, tanto o tipo financiado *mortgage* quanto a própria *own* apresentaram sinais negativos e esperados uma vez que o tipo alugado *rent*, que é a referência, é a mais arriscada pois além do custo de moradia o devedor não tem posse e, portanto, menos ativos próprios. Quanto ao uso dos recursos não existe teoria prévia, mas importante notar que nas finalidades consolidação de dívidas *debt_consolidation*, gastos com a casa *house*, outros *other* e mudança *moving*, o sinal é negativo, ou seja, quando os empréstimos têm essas finalidades, a recuperação em caso de inadimplência é maior. O contrário foi observado na variável *home_improvement*, que apresentou sinal positivo, ou seja, uma maior perda.

Tabela 6
Método Tradicional – Árvore de Decisão

Métricas	100%	10%	0,5%
Variável - 1º nó	ead_loan	annual_inc	ead_loan
Variável - 2º nível - nó A	annual_inc	ead_loan	int_rate
Variável - 2º nível - nó B	annual_inc	ead_loan	dti
Variável - 3º nível - nó A	ead_loan	ead_loan	ead_loan
Variável - 3º nível - nó B	home_ownership: mortgage	loan_amnt	-
Variável - 3º nível - nó C	fico_range_low	emp_length	-
Variável - 3º nível - nó D	ead_loan	ead	loan_amnt
Núm. observações	97147	9714	485

A tabela indica os resultados do modelo de árvore de decisão para a estimação do LGD, limitando o máximo número de níveis à 10 e uma quantidade mínima de observações em cada nó folha a 2,5% da amostra de desenvolvimento. **Variável e Nó:** indica qual a variável foi selecionada para segmentação do respectivo nível da árvore, do 1º ao 3º nível. Elas são detalhadas na tabela 1. **Núm. Observações:** indica a quantidade de observações utilizadas no desenvolvimento da árvore, representando respectivamente o percentual da amostra utilizada no cabeçalho da tabela.

Finalmente, apesar de significante o efeito esperado de algumas variáveis não foi o previsto na literatura e indicado na tabela 2. A classificação de crédito *grade* mostrou que quanto melhor o valor maior o LGD, o mesmo observado na variável *mths_since_last_delinq_null* que indica que o devedor não teve nenhum atraso no histórico no momento da concessão. Nas análises bi-variadas o mesmo foi observado e isso indica que a empresa é bem menos conservadora na determinação do valor do

empréstimo quando os clientes apresentam histórico de crédito muito bom, o que piora a recuperação de crédito nesses casos e consequentemente o LGD.

Tabela 7
Método Tradicional – Regressão Linear

Amostra	**100%**		**10%**		**0,5%**	
Variáveis	**Coefic.**	**P-valor**	**Coefic.**	**P-valor**	**Coefic.**	**P-valor**
constante	0,8692	0	0,8968	0	0,8825	0
term	0,0021	0,026				
ead_loan	0,0356	0				
emp_lenght	-0,001	0	-0,0013	0		
dti	-0,0001	0,009	-0,0004	0,022		
delinq_2yrs	-0,0017	0				
mths_since_last_delinq_null	0,006	0	0,0068	0,012		
pub_rec	0,0029	0				
grade: B	0,0058	0				
grade: C	0,0046	0				
grade: D	0,0025	0,033				
home_ownership: mortgage	-0,0104	0	-0,0072	0,013		
home_ownership: own	-0,0093	0	-0,0126	0,006		
purpose: debt_consolidation	-0,0035	0				
purpose: house	-0,0122	0,06				
purpose: other	-0,0057	0,006				
purpose: home_improvement					0,051	0,021
purpose: moving			-0,0306	0,037	-0,1558	0,013
ead			-0,0097	0		
ead_inc			0,1223	0		
pub_rec			0,0043	0,036		
Núm. observações	97147		9714		485	
R^2	0,008		0,01		0,024	
R^2 Ajustado	0,008		0,009		0,02	

A tabela apresenta os resultados do modelo de regressão linear simples aplicado sobre 3 amostras de desenvolvimento de diferentes tamanhos para a estimação do LGD. **Variáveis** referem às variáveis finais selecionadas pelos modelos. **Núm. Observações** indica o tamanho da amostra utilizada no desenvolvimento. **R^2 e R^2 Ajustados** indicam a qualidade de ajuste do modelo. **Coef.** refere-se ao coeficiente e respectivo sinal do parâmetro da variável. **P-valor** indica o nível de significância do parâmetro.

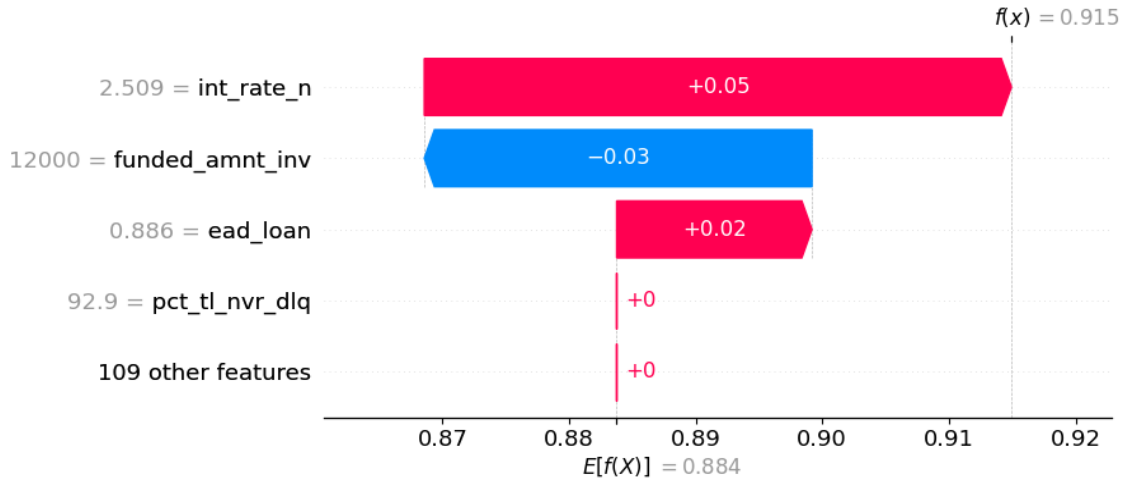
O contrário ocorreu com as variáveis "*dti*" que indica o endividamento no cliente no momento da concessão e com a variável "*delinq_2yrs"* que indica o número de atrasos nos últimos 2 anos, assim apesar de esperado um valor positivo para os parâmetros observamos sinais negativos e isso pode ser também produto da política de crédito, que é provavelmente muito restritiva para devedores nessas condições e, portanto, na ocorrência da inadimplência a recuperação é melhor. O mesmo efeito foi observado na variável "*ead*" que indica o montante devido no momento da inadimplência, o que reforça a efetividade da política de limites adotada pela empresa.

Zhou *et al*. (2018) observaram os mesmos resultados esperados reportados aqui para as variáveis *term*, *emp_lenght* e *home.* As variáveis *dti* e *delinq_2yrs* também apresentaram os sinais esperados, diferente dos resultados desse estudo, porém isso pode decorrer da amostragem realizada pelos autores, que excluiu uma parte relevante dos dados para o desenvolvimento do modelo.

Para a avaliação das variáveis relevantes dos modelos de AutoML foi utilizado o método Kernnel Shap. A figura 3 a seguir ajuda a entender o valor de Shap para uma única observação extraída dos dados de treinamento. A partir do valor médio de 0,884 do LGD observado na amostra de treinamento o gráfico decompõe os pesos das variáveis que mais contribuem para o valor estimado de 0,915 como desvio do valor médio, destacando a parte positiva sobre o LGD determinados pelo *ead_loan* (+0,02) e *int_rate* (+0,05) detalhados na tabela 2 e a parte negativa, ou seja, reduzindo o LGD, influenciado pelo *funded_amnt_inv* (-0,03) que se refere ao valor não amortizado do investidor observado na data. O valor de Shap para um conjunto de variáveis refere-se à média dos valores individuais em termos absolutos, assim, quanto maior mais importante a variável. Na observação abaixo a variável mais relevante é o *int_rate*.

A tabela 8 apresenta os valores de Shap, ou seja, a importância das variáveis tomadas pela média dos valores absolutos observados (Gianfagna e Di Cecco, 2021) considerando 500 simulações em cada modelo de AutoML, de onde foram selecionadas as 10 mais relevantes. Foi tomado o valor médio de todas as 10 simulações de cada cenário. Como a variável de interesse, o LGD, apresenta uma média amostral de aproximadamente 0,88 os valores apresentados foram multiplicados por 100 para facilitar a leitura e comparação. Não é possível fazer uma comparação completa, pois as variáveis *ead*, *ead_inc* e *ead_loan* não foram utilizadas no treinamento do modelo de AutoML.

Figura 3
Valores de Shap calculados para 1 observação aleatória



O eixo X apresenta o valor do LGD da amostra de desenvolvimento. Nas linhas são indicadas as variáveis por ordem de valor de Shap em termos absolutos que contribuem para o valor estimado da observação selecionada indicado no alto à direita na figura.

Tabela 8
Valores de Shap – AutoML

Amostragem	**100%**		**10%**		**0,5%**		**Qtd**	**Média**
Tempo processamento	**1h**	**3h**	**1h**	**3h**	**1h**	**3h**		
Variáveis								
Loan_amnt	5,5	5,32	5,24	7,97	5,65	5,43	6	5,85
Funded_amnt_inv	3,91	4,55	3,82	8,23	5,56	5,58	6	5,28
Term	2,32	2,44	2,2	5,01	3,12	3,19	6	3,05
Int_rate	1,26	1,39	1,29	2,92	1,93	1,81	6	1,77
Installment	0,89	1,01	0,71	1,89	1,18	1,12	6	1,13
Sub_grade	0,61	0,69	0,53	1,21	0,69	0,80	6	0,76
Total_pymnt	0,51	0,53	0,43	1,02	0,71	0,64	6	0,64
Emp_lenght	0,32	0,47	0,37	0,59	0,58	0,49	6	0,47
Purpose: Vacation	1,55	1,69			1,68	1,72	4	1,66
Annual_inc			0,35	0,73	0,51	0,52	4	0,53
Total_rec_prncp		0,31	0,29				2	0,3
Earliest_cl_line				0,60			1	0,6
Verification	0,39						1	0,39

A tabela apresenta os valores de Shap para 500 simulações do modelo de AutoML considerando todos os cenários de teste em termos de amostragem (100%, 10% e 5%) e tempo de processamento (1hora, 3 horas). **Variáveis:** indica as variáveis mais relevantes em termos do do valor de Shap, considerando as 10 mais relevantes. **Qtd:** Indica em quantos modelos as variáveis foram selecionadas. **Média**: mostra o valor médio considerando os modelos onde a variável foi relevante.

Apesar da utilização de mais de 100 variáveis no treinamento do modelo de AutoML é clara a convergência para variáveis relevantes na literatura sobre o tema LGD. Fixando a análise nas 8 principais constata-se que as duas primeiras, *loan_amnt* e *funded_amnt_inv* referem-se ao volume emprestado não amortizado, tanto com relação ao valor nominal do empréstimo quanto o proporcional ao recurso captado com investidores e cabe destacar também elas são muito semelhantes, tanto que a correlação de Pearson entre ambas é de 0,9999. A primeira também foi relevante no modelo de árvore de decisão. A variável seguinte *term,* relativa ao prazo do empréstimo foi relevante na análise de regressão e a próxima, taxa de juros, representada por *int_rate* foi relevante na árvore de decisão.

A variável *installment* se refere ao valor da parcela e ela é normalmente proporcional ao valor do empréstimo, tanto que a correlação de Pearson entre ela e *loan_amnt* é 0,938, o que explica sua relevância para o modelo. *Sub_grade* nada mais é que uma maior segmentação da variável *grade,* ou seja, a classificação de risco do empréstimo no momento da contratação e que foi relevante na análise de regressão.

A variável seguinte *total_pymnt* refere-se ao total amortizado no momento da inadimplência e ela é relevante conceitualmente pois quanto menor o seu valor, maior o *ead,* ou a exposição no momento da inadimplência, e contatamos que tanto o *ead* como suas versões relativas foram extremamente relevantes nos modelos tradicionais desenvolvidos neste artigo. Finalmente o tempo no emprego do devedor *emp_lenght* foi relevante tanto na análise de regressão quanto na árvore de decisão. Além da grande convergência observada neste artigo, importante destacar que das 8 principais variáveis desta análise, 5 também estiveram entre as 8 mais significantes na análise de regressão de Zhou *et al*. (2018) e foram elas *loan_amnt, term, int_rate, grade* e *emp_lenght*.

Esses resultados mostram, mesmo no contexto de uso de uma poderosa ferramenta de análise como o AutoML é fundamental procurar na teoria o referencial relevante para o problema analisado. Fica evidente que a capacidade do algoritmo em identificar adequadamente as complexas relações entre as variáveis explica em grande parte o sucesso do resultado observado aqui. A desmistificação do ML como uma caixa preta e a robustez dos resultados em termos de performance de todas as simulações deixam clara a importância e a oportunidade do uso dessas metodologias nos modelos de parâmetros de risco de crédito.

6. Conclusão

O sucesso do uso de métodos de Machine Learning nas últimas décadas provocou uma revolução em vários negócios. Uma evolução recente nesse ramo foi o desenvolvimento de algoritmos de Aprendizado Automático de Máquina (AutoML), cujo objetivo é substituir todas as etapas do desenvolvimento de um modelo de Machine Learning, extremamente custosas em termos de consumo de tempo humano, por um algoritmo que procura otimizar, além da performance, o tempo e custo de processamento.

A aplicação dessas metodologias no mercado financeiro é desafiadora, principalmente no que se refere aos parâmetros de perda esperada devido à ênfase dos reguladores em explicação e comparação no lugar de acurácia, o que limita o campo de aplicação desses métodos, muitas vezes chamados de "caixa-preta" por não permitirem um claro entendimento dos fatores que determinam as estimações, bem como suas relações econômicas. Em função disso, a literatura sobre o tema é escassa, principalmente em se tratando de modelos de previsão do LGD. Dessa forma, este artigo preenche uma importante lacuna na área de risco de crédito, sendo o primeiro a explorar um algoritmo de AutoML na estimação um parâmetro de perda específico e avaliando, além da diferença da performance, os principais determinantes do resultado em termos de variáveis explicativas, principal crítica dos reguladores.

Com relação à performance, o AutoML mostrou-se muito superior aos métodos tradicionais tanto em termos de acurácia quanto ordenação. O AutoML reduziu o erro em mais de 50% em todas as métricas avaliadas, sendo robusto até mesmo na menor amostra de treino. Considerando as métricas de ordenação, a diferença é ainda significativa, em alguns casos melhorando cerca de 10 vezes o indicador. Com relação ao tamanho da amostra, de fato a performance do AutoML cai consideravelmente com a sua redução, o que pode ser um fator limitante para algumas aplicações quando não há um número relevante de observações, mas mesmo nos cenários mais críticos a performance ainda foi muito superior à metodologia tradicional. Finalmente não foi observada uma variação relevante de performance do modelo quando o tempo de processamento foi reduzido de 3 horas para apenas 1 hora, o que é relevante pois permite o desenvolvimento de um maior número de simulações na mesma máquina sem perda relevante qualidade do modelo.

Apesar da utilização de mais de 100 variáveis em todos os cenários de treino do AutoML, a relevância medida pelo valor de Shap indicou que as variáveis mais

importantes foram em sua maioria as mesmas selecionadas pelo método tradicional, o que reforça a consistência e segurança dos resultados obtidos, bem como a capacidade da metodologia em identificar relações não lineares entre as variáveis, otimizando o seu uso. Esse conjunto de resultados indicam uma vantagem clara do uso de AutoML em termos de performance, custo e uma menor necessidade de especialização dos envolvidos no processo, uma vez que a maioria das etapas do desenvolvimento são realizados pelo algoritmo.

No que se refere às limitações desse artigo e oportunidade de estudos futuros quatro pontos podem ser abordados. Primeiro, para manter a comparabilidade entre os métodos, uma clara vantagem do AutoML que não foi explorada se refere ao processo de limpeza e tratamento automático das variáveis, o que em geral toma pelo menos metade do tempo de desenvolvimento de um modelo. Segundo, em estudos de ML normalmente há também um processo prévio de seleção de variáveis cujo objetivo é retirar eventuais ruídos e facilitar a implantação e monitoramento dos modelos, fundamental para a indústria financeira, mas isso também estava fora do escopo desse artigo. Terceiro, além do valor de Shap outras formas de avaliação da relevância das variáveis e seus impactos podem ser exploradas, pois uma mudança estrutural nas operações e consequentemente nos dados pode acarretar uma rápida deterioração do modelo, principalmente quando a variável de interesse é observada muito tempo à frente (Bazarbash, 2019). A interpretabilidade do modelo deve ser acessível a todos os agentes, desde os responsáveis pela gestão de crédito até os clientes, pois a definição de preço e a aprovação ou não de novos empréstimos possui importante impacto econômico (Alonso e Carbó, 2020; Escalante, 2021; Israel *et al.*, 2020) e regulatório (EBA, 2021). Quarto, a distribuição e complexidade do LGD varia muito em função de produto e indústria, o que não garante os mesmos resultados obtidos aqui em outras aplicações, além disso os demais parâmetros de perda PD e EAD também podem ser objeto de modelagem utilizando AutoML. Todas essas possíveis aplicações podem contar com um número relevante de novos algoritmos e técnicas que têm sido desenvolvidos nos últimos anos.

Finalmente, como a melhor performance dos modelos de ML e AutoML são obtidos ao custo de maior complexidade e menor compreensão das relações entre as variáveis, para viabilizar o seu uso pelas instituições EBA (2021) traz algumas recomendações: 1) evitar o uso de variáveis que não melhorem a performance geral do modelo; 2) evitar o uso de dados desestruturados quando estruturados similares estão

disponíveis; 3) em termos de escolha de algoritmo, usar preferencialmente aqueles mais simples quando a performance não é muito diferente. Dessa forma, os modelos podem ser comunicados de forma mais simples para os gestores de risco e reguladores.

Existe ainda um amplo campo de pesquisa na área de ML e AutoML em crédito e debater esses resultados e alternativas é a forma mais produtiva de aumentar o uso desses métodos, melhorando as estimativas e garantindo uma melhor gestão de risco e resultados das instituições financeiras. A recente evolução nas técnicas de avaliação de modelos permite o melhor entendimento do seu funcionamento, viabilizando também o desenvolvimento de modelos aderentes aos requisitos do Acordo de Basiléia. Esse artigo dá um primeiro passo nesse sentido.

Referências

Alonso, A., Carbó, J.M., (2020). Machine Learning in Credit Risk: Measuring the Dilema Between Prediction and Supervisory Cost, *Documento de Trabajo nº 2032,* Banco de España.

Bazarbash, M. (2019). Fintech in Financial Inclusion: Machine Learning Applications in Assessing Credit Risk, *WP/19/109,* International Monetary Fund.

BIS - Bank for International Settlements (2004). Basel II: International Convergence of Capital Measurement and Capital Standards: A Revised Framework, Basel: BIS, June.

BIS - Bank for International Settlements (2005). Studies on the Validation of Internal Rating Systems, Basel: BIS, May.

Breiman, L (2001). Statistical Modeling: The Two Cultures, *Statistical Science* **16** (3): 199-215.

Dermine, J., Neto de Carvalho, C. (2006). Bank Loan Losses-Given-Default: A Case Study, *Journal of Banking and Finance* 30:1219-1243.

EBA – European Bank Authority (2021). Discussion Paper on Machine Learning for IRB Models, *EBA/DP/2021/04*.

Escalante, H.J. (2021). Automated Machine Learning—A brief review at the end of the early years, *in* N. Pillay, R. Qu (eds), *Automated Design of Machine Learning and Search Algorithms*. Natural Computing Series. Springer Nature Switzerland.

Feurer, M., Eggensperger, K., Falkner, S., Lindauer, M., Lindauer, M., Hutter, F. (2022). Auto-Sklearn 2.0: Hands-Free AutoML via Meta-Learning, *Journal of Machine Learning Research* **23**: 1-61.

Feurer, M., Hutter, F. (2019). Hyperparameter Optimization *in* F. Hutter, L. Kotthoff, J. Vanschoren (eds), *Automated Machine Learning: Methods, Systems, Challenges. The Springer Series on Challenges in Machine Learning*. Switzerland, pp. 3-33.

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J. T., Blum, M., Hutter, F. (2019). Auto-sklearn: Efficient and Robust Automated Machine Learning *in* F. Hutter, L. Kotthoff, J. Vanschoren (eds), *Automated Machine Learning: Methods, Systems, Challenges. The Springer Series on Challenges in Machine Learning*. Switzerland, pp. 113-134.

Gianfagna, L., Di Cecco, A. (2021). *Explainable AI with Python*, Springer Nature Switzerland.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., Altman, D. G. (2016). Statistical Tests, P values, Confidence Intervals, and power: a Guide to Misinterpretations, *European Journal of Epidemiology* 31: 337-350.

Guégan, D., Hassani, B. (2018). Regulatory Learning: How to Supervise Machine Learning Models? An Application to Credit Scoring, *The Journal of Finance and Data Science*, **4**: 157-171.

Israel, R., Kelly, B., Moskowitz, T. (2020). Can Machines "Learn" Finance? *Journal of Investment Management*, **18**(2): 23-36.

Lima, J. C. C. O. (2008). A Importância de Conhecer a Perda Esperada para Fins de Gerenciamento do Risco de Crédito, *Revista do BNDES* 15: 271-302.

Loterman, G., Brown, I., Martens, D., Mues, C. e Baesens, B. (2012). Benchmarking Regression Algorithms for Loss Given Default Modeling, *International Journal of Forecasting* **28**(1): 161-170.

Lundberg, S. M., Lee, S. (2017). A Unified Approach to Interpreting Model Predictions, *in 31st Conference on Neural Information Processing Systems*, Long Beach, CA, USA.

Papoušková, M., Hajek, P. (2019). Modelling Loss Given Default in Peer-to-Peer Lending Using Random Forest. *in* I. Czarnowski, R. J. Howlett, L. C. Jain (eds), *Intelligent Decision Technologies 2019. Smart Innovation, Systems and Technologies, vol 142*. Springer, Singapore, pp. 133-141.

Qi, M., Zhao, X. (2011). Comparison of Modeling Methods for Loss Given Default, *Journal of Banking and Finance* **35**(11): 2842-2855.

Scandizzo, S. (2016). *The Validation of Risk Models: A Handbook for Practitioners*, Palgrave Macmillan.

Schuermann, T. (2004). What Do We Know about Loss Given Default? *in* D. Shimko (ed), *Credit Risk Models and Management*, 2. ed., Risk Books, London, p. 249-274.

Silva, A. C. M. da, Marins, J. T. M. e Neves, M. B. E. (2009). Loss Given Default: Um Estudo sobre Perdas em Operações Prefixadas no Mercado Brasileiro, *Working Paper 193*, Banco Central do Brasil.

Tuggener, L., Amirian, N., Rombach, K., Lörwald, S., Varlet, A., Westermann, C., Stadelmann, T., (2019). Automated Machine Learning in Practice: State of the Art and Recent Results. *in 6th Swiss Conference in Data Science (SDS)*, Bern, Switzerland.

Yao, X., Crook, J., Andreeva, G. (2015). Support Vector regression for Given Default Modelling, *European Journal of Operational Research* **240**(2): 528-538.

Yashkir, O., Yashkir, Y. (2013). Loss Given Default Modeling: A Comparative Analysis, *Journal of Risk Model Validation* **7**(1): 25-59.

Zhou, G., Zhang, Y. e Luo, S. (2018). P2P Network Lending, Loss Given Default and Credit Risks, *Sustainability* **10**(4): 1010.

Zöller, M., Huber, M. F. (2021) Benchmark and Survey of Automated Machine Learning Frameworks, *Journal of Artificial Intelligence Research* **70**: 409-472.