Análise de Risco de Crédito para Gestoras de Vendedores Porta-a-Porta usando Aprendizado de Máquina ⋆

Ricardo Zorzal Davila [*,**] **Francisco de Assis Boldt** [*] **Filipe Mutz** [*]

[*] *Programa de Pós-Graduação em Computação Aplicada, Instituto Federal do Espírito Santo (IFES) - Campus Serra, ES (emails: {franciscoa, filipe.mutz}@ifes.edu.br).*
[**] *Máximus Soluções, ES (e-mail: ricardozorzaldavila@gmail.com).*

Abstract: Credit risk analysis is fundamental for small and medium companies given that defaults cause significant impacts in revenue. Banks and other financial institutions have access to several information regarding the financial wellbeing of clients for credit analysis. On the other hand, companies from other areas need to estimate credit risk using demographic information and the history of interations with the company. This work studies the problem of credit risk analysis for companies that manage door-to-door salesmen. These companies have specific characteristics that differenciate them from businesses of other niches. These specificities influence the process of credit risk analysis. A dataset was built using information from partner companies and this dataset was used to train several machine learning algorithms in the task of default prediction. Experiments showed that the logistic regression classifiers achieves the same performance as nonlinear classifiers (e.g., neural networks, XGBoost, and ensembles), while being less prone to overfitting and being more interpretable.

Resumo: A análise de risco de crédito é de fundamental importância para pequenas e médias empresas uma vez que o não pagamento de compras gera impactos significativos nos rendimentos. Bancos e financeiras possuem acesso à diversas informações relacionadas à saúde financeira dos clientes para análise de crédito. Já empresas de outros ramos precisam calcular o risco de crédito usando informações demográficas de clientes e seu histórico de interações com a empresa. Este trabalho estuda o problema de análise de risco de crédito para empresas gestoras de vendedores porta-a-porta. Estas empresas possuem características específicas que as diferenciam de negócios de outros segmentos. Estas especificidades influenciam o processo de análise de risco de crédito. Foi construída uma base de dados usando informações de empresas parceiras e esta base foi usada para treinar diversos algoritmos de aprendizado de máquina na tarefa de predição de não pagamento. Experimentos mostraram que o classificador linear *logistic regression* é capaz de alcançar a mesma performance de classificadores não lineares (e.g., redes neurais, XGBoost e *ensembles*), sendo menos suscetível à *overfitting* e mais interpretável.

Keywords: Credit Risk Analysis; Credit Scoring; Artificial Intelligence; Machine Learning; XGBoost; Neural Networks; Logistic Regression.

Palavras-chaves: Análise de Risco de Crédito; Pontuação de Crédito; Inteligência Artificial; Aprendizado de Máquina; XGBoost; Redes Neurais; *Logistic Regression*.

1. INTRODUÇÃO

O setor de crédito é de considerável importância econômica. Em 2019, o crédito ao consumidor em aberto nos EUA era de aproximadamente US$ 4.11 trilhões (Federal Reserve System, 2021; Statista, 2021). No Brasil, as estatísticas são similares. Em fevereiro de 2021, o saldo total de empréstimos e financiamentos ao setor não financeiro no Brasil era de aproximadamente R$ 4.3 trilhões (Banco Central do Brasil, 2021).

O número de empréstimos e a taxa de inadimplência no pagamento de empréstimos é influenciada pela situação social, política e econômica dos países. De 2011 a 2021, a taxa de inadimplência de carteiras de crédito para pessoas físicas oscilou entre 3% e 4,6%, alcançando valor máximo de 5,5% no Brasil. Mesmo com a política de juros em caso de atraso de pagamento, empresas que optam por dar crédito aos clientes podem sofrer impactos financeiros devido à inadimplência. Estes impactos são particularmente significativos em empresas de pequeno e médio porte. Buscando minimizar estes impactos, é comum o uso de

⋆ Agradecemos à FAPES e a CAPES pelo apoio financeiro dado por meio do PDPG (Parcerias Estratégicas nos Estados da CAPES) (PROCESSO: 2021-2S6CD, TO/nº FAPES: 132/2021). Também agradecemos ao Propós (Programa Institucional de Apoio à Pós-graduação Stricto Sensu) do IFES pela apoio financeiro. Filipe Mutz agradece ao Instituto Federal do Espírito Santo (IFES) por incentivar sua pesquisa via o Programa Pesquisador de Produtividade (PPP) - portaria n. 1072 de 21 de maio de 2020.

técnicas de análise de risco de crédito para avaliar se é mais vantajoso para a empresa aprovar um empréstimo ou negá-lo devido ao risco de inadimplência. Bancos e financeiras possuem mecanismos para obter informações sobre salário, patrimônio, histórico de compras, pagamentos e dívidas dos clientes. Já empresas de outros ramos, precisam realizar a análise de risco de crédito utilizando apenas informações demográficas (e.g., gênero, idade e local de moradia), e o histórico de transações com a empresa (e.g., compras e pagamentos). Por isto, a taxa de ruído na análise é significativamente maior nestas empresas em comparação com bancos e financeiras.

Um tipo de negócio muito comum em comunidades de baixa renda são os vendedores porta-a-porta. Estes vendedores visitam as casas dos clientes para apresentar novos produtos, realizar vendas e receber pagamentos. As vendas neste modelo são comumente feitas à prazo e muitos dos clientes atendidos não possuem crédito disponível via outros meios. Recentemente, foram estabelecidas empresas especializadas que agregam e gerenciam estes tipos de vendedores. Estas empresas adicionam profissionalismo e impessoalidade aos negócios com objetivo de maximizar os lucros. Elas também armazenam informações sobre clientes e vendedores com objetivo de viabilizar o desenvolvimento de sistemas para otimização de rotas e recomendação de produtos.

Tradicionalmente, a decisão de realizar ou não a venda à prazo na venda porta-a-porta é feita pelos vendedores com base no histórico de transações e em fatores subjetivos como a relação pessoal com os clientes. Este tipo de análise não-objetiva pode ser sub-ótima do ponto de vista da lucratividade do negócio e, além disso, pode carregar preconceitos e visões distorcidas do vendedor. Assim, ferramentas capazes de automatizar o processo podem trazer benefícios tanto para as empresas quanto para os clientes.

Portanto, este trabalho apresenta e avalia um sistema para análise de risco de crédito para empresas gerenciadoras de vendedores porta-a-porta. O sistema utiliza modelos de aprendizado de máquina para classificar se uma venda é segura ou se existe uma probabilidade alta de calote. Até onde sabemos, este é o primeiro trabalho com foco neste tipo de negócio. Ele possui características únicas que o diferencia de instituições financeiras como o público-alvo complementar e informações limitadas sobre o cliente, suas posses, rendimentos e saúde financeira. Estas diferenças fazem com que adaptações nos métodos da literatura sejam necessárias, além de avaliações experimentais para verificar que técnicas e configurações são mais adequadas ao contexto.

Foi construída uma base de dados junto a empresas parceiras contendo dados de cadastro de clientes e seu histórico de compras, agendamentos de visitas e pagamentos de parcelas. Para cada compra, foi identificado se o cliente realizou os pagamentos no prazo ou se houve inadimplência. Estes dados foram utilizados para treinar classificadores individuais e *ensembles* na tarefa de predizer se as compras serão pagas no tempo acordado ou não. Os classificadores foram avaliados usando o método de *K-fold cross-validation* usando as métricas acurácia, precisão, revocação e F1-*score*. Além disso, foram realizados experimentos qualitativos com objetivo de interpretar se as regras aprendidas pelos modelos não são padrões espúrios dados. Para isto, as importâncias dos classificadores baseados em árvore de decisão e os coeficientes de um classificador linear foram analisados.

Resultados experimentais mostraram que todos os métodos individuais, exceto o Gaussian Naive Bayes (GNB) (Bishop, 2006), são capazes de predizer o risco de atrasos no pagamento com acurácia de aproximadamente 80% e que o uso de *ensembles* não leva à ganhos significativos de performance nesta base de dados. A análise das importâncias e coeficientes das características mostrou que os métodos consideram características que fazem sentido na tomada de decisão.

2. TRABALHOS RELACIONADOS

O estudo de ferramentas para análise automática do risco de crédito não é novo. Há mais de vinte anos, Hand e Henley fizeram um *survey* sumarizando algoritmos para solução do problema e os desafios encontrados na implantação destes sistemas (Hand and Henley, 1997). Posteriormente, Lessmann et al. (2015) realizaram uma revisão atualizada até 2015 comparando métodos baseados em aprendizado de máquina para análise de risco (Lessmann et al., 2015). Louzada et al. (2016) fizeram uma revisão sistemática de literatura sobre pontuação de crédito e categorizaram os métodos usados para automatizar a análise (Louzada et al., 2016). Dastile et al. (2020) realizaram uma revisão sistemática da literatura recente sobre o tema.

Trabalhos recentes têm usado aprendizado de máquina, em particular ensembles de árvores de decisão e *deep learning*, para construir preditores de risco de crédito usando dados históricos (Xia et al., 2017; Luo et al., 2017; He et al., 2018; Pławiak et al., 2020). A maioria destes trabalhos se concentram em bancos e financeiras.

Xia et al. (2017) aplicaram *eXtreme gradient boosting* (XGB) para pontuação de crédito usando dados de bancos. Experimentos mostraram que esta técnica levou à resultados melhores que métodos tradicionais como busca em grid, busca aleatória e busca manual. Buscando interpretar as regras aprendidas, os autores proveem os pesos das características e os nós das árvores de decisão que compõe o *ensemble*.

Luo et al. (2017) compararam a performance de *deep belief networks* (DBN), um tipo de rede neural com múltiplas camadas construído empilhando redes neurais do tipo *restricted Boltzmann machines*, com classificadores tradicionais, e.g. regressão logística, SVMs e redes neurais *multilayer perceptron*. Nos experimentos realizados, as DBNs alcançaram performance superior.

Pławiak et al. (2020) propuseram o *Deep Genetic Hierarchical Network of Learners*, um método para classificação de risco de crédito construído usando uma variedade de classificadores, de técnicas de treinamento, de pré-processamentos e de métodos de otimização de parâmetros. Usando esta sofisticada arquitetura de métodos, os autores alcançaram resultados estado-da-arte na base de dados de aprovação de crédito *Statlog German*, uma das bases mais usadas na literatura (infelizmente as informações da base são de um contexto diferente deste trabalho).

Embora trabalhos recentes usem predominantemente métodos de aprendizado de máquina, é digno de nota que Ben-David and Frank (2009) compararam sistemas especialistas e métodos de aprendizado de máquina na predição de pontuação de crédito e não observaram diferença significativa entre as técnicas quando a tarefa foi modelada como um problema de classificação. Quando ela foi modelada como regressão, os métodos de aprendizado de máquina alcançaram ligeira vantagem. Vale enfatizar que os métodos foram comparados em termos de performance preditiva e critérios importantes como a clareza das regras usadas na decisão não foram considerados.

3. BASE DE DADOS

Para o desenvolvimento do sistema proposto neste trabalho foi construída uma base de dados junto a empresas parceiras com informações de cadastro de clientes e as interações dos clientes com a empresa, e.g., compra de produtos, pagamentos de parcelas e agendamentos desmarcados. Para garantir a privacidade dos clientes, a base foi anonimizada e todos os dados que poderiam ser usados para identificar o cliente foram removidos.

A tarefa dos modelos é de classificação e consiste em predizer se uma compra será paga no prazo ou se acontecerão atrasos. Para isto, ele recebe como entrada as informações de cadastro dos clientes e da compra, além do histórico de compras, pagamentos e agendamentos até a data da compra. A saída é binária e possui valor 1 se a compra foi paga no prazo e valor 0 se a compra resultou em atraso. As empresas consideram que uma compra foi paga no prazo se ela foi paga em até dois meses. Desta forma, a compra é considerada atrasada se o prazo de pagamento for superior a dois meses. Compras realizadas a menos de dois meses em relação à data de criação da base de dados foram descartadas já que não era possível definir se ela seria paga no prazo ou não.

Para evitar que os modelos memorizassem dados dos clientes, foi selecionada uma compra aleatória por cliente e apenas os dados relacionados à esta compra foram mantidos na base de dados (as demais foram descartadas). Com isto, existe apenas uma amostra por cliente o que previne que dados do mesmo cliente estejam no conjunto de treino e teste. Todas as informações posteriores à data da compra são descartadas para garantir que os modelos não tenham acesso à dados futuros.

Considerar apenas uma compra por cliente pode parecer um desperdício de informações, contudo como a base de dados possui um grande número de amostras, a subamostragem não foi prejudicial. Foram coletados dados de 5 empresas, totalizando 27.230 clientes e, portanto, de compras. Um total de 16.187 (59%) destas compras foram pagas no prazo, enquanto 11.043 (41%) foram pagas com atraso ou resultaram em calote.

Dentre as características usadas como entrada para os modelos podemos citar o gênero do cliente, o saldo até o momento com a empresa (diferença entre o total comprado e o total pago), o valor da compra sendo realizada, meio de pagamento utilizado, valores e datas de compras e pagamentos anteriores, datas de agendamentos de visitas e indicadores se eles foram cancelados ou não.

As características da base de dados foram pré-processadas como descrito a seguir. Dados categóricos foram codificados usando o formato *one-hot* (Bishop, 2006). Campos que representam datas foram transformadas no número aproximado de anos (valor de ponto flutuante) em relação à data de download da base de dados. Para calcular este valor, foram contados quantos dias haviam se passado entre as datas e o valor foi dividido por 365. Quanto maior o resultado, mais antiga é a data. Para equalizar o número de agendamentos, compras e pagamentos até a data da compra, foi utilizada a técnica de *padding* na qual as sequências com menos elementos que um valor máximo pré-definido são preenchidas com valores nulos ou inválidos. O número máximo de pagamentos e compras foi de 40 cada, enquanto o número máximo de agendamentos foi de 100.

4. EXPERIMENTOS

Esta seção descreve os experimentos realizados para avaliar a viabilidade de usar modelos de aprendizado de máquina para predizer se compras seriam pagas no prazo ou não. Cada subseção representa um experimento e discute a metodologia utilizada e os resultados alcançados. Os dados e o código utilizado nos experimentos estão disponíveis em `https://github.com/rzorzal/mestrado-financial-classification`.

4.1 Comparação de Modelos de Aprendizado de Máquina

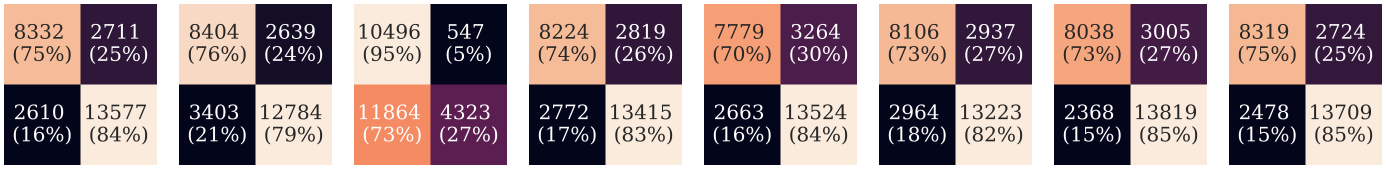
O primeiro experimento teve como objetivo avaliar se os modelos de aprendizado de máquina são capazes de predizer o resultado das compras e identificar qual modelo tem a melhor performance preditiva. Os modelos considerados foram *Extreme Gradient Boosting* (XGB) (Chen et al., 2015), *K-Nearest Neighbours* (KNN), *Decision Tree* (DT), *Random Forest* (RF), Redes Neurais do tipo *Multilayer Perceptron* (MLP), *AdaBoost* (AB) e *Gaussian Naive-Bayes* (GNB) (Bishop, 2006). Para avaliar se um classificador linear poderia resolver o problema, também foi utilizado o método *Logistic Regression* (LR) (Bishop, 2006). Outros classificadores lineares poderiam ter sido selecionados e não houve motivação particular na escolha do LR.

A performance dos modelos foi avaliada usando o método de *K-fold cross-validation* com 10 *folds* e as métricas consideradas foram acurácia, precisão, revocação e f1-*score* (Bishop, 2006). A acurácia foi calculada como a média entre *folds*. Já para calcular os valores de precisão, revocação e f1-*score*, uma matriz de confusão foi produzida para cada *fold* e a média (macro) entre as métricas para cada classe foi calculada. Por fim, foram calculadas estatísticas (média e desvio padrão) dos valores entre *folds*.

O código-fonte dos experimentos foi desenvolvido a linguagem Python (versão 3.8.5) e a implementação dos modelos, métricas e do método *K-fold cross-validation* vieram da biblioteca scikit-learn (versão 0.24) Pedregosa et al. (2011). A implementação do método XGBoost veio da biblioteca com mesmo nome (versão 1.4.1) (`https://xgboost.readthedocs.io/en/latest/index.html`). Foram usados os valores padrão de hiperparâmetros das bibliotecas, com exceção do número de épocas da MLP que foi modificado de 200 para 1000.

Tabela 1. Comparação de modelos de aprendizado de máquina na análise de risco de crédito.

Modelo	Acurácia	Precisão	Revocação	F1-*Score*
XGB	**0.81 ± 0.007**	**0.80 ± 0.006**	**0.80 ± 0.007**	**0.80 ± 0.008**
KNN	0.78 ± 0.008	0.78 ± 0.011	0.77 ± 0.009	0.77 ± 0.010
DT	0.78 ± 0.008	0.77 ± 0.007	0.78 ± 0.008	0.77 ± 0.008
RF	0.80 ± 0.009	**0.80 ± 0.009**	0.79 ± 0.009	0.79 ± 0.009
MLP	0.78 ± 0.020	0.78 ± 0.012	0.78 ± 0.014	0.77 ± 0.017
AB	0.80 ± 0.008	**0.80 ± 0.008**	**0.80 ± 0.009**	**0.80 ± 0.009**
GNB	0.54 ± 0.009	0.68 ± 0.007	0.61 ± 0.007	0.52 ± 0.010
LR	0.79 ± 0.007	0.79 ± 0.006	0.79 ± 0.007	0.79 ± 0.007



(a) AB (b) DT (c) GNB (d) LR (e) KNN (f) MLP (g) RF (h) XGB

Figura 1. Matrizes de confusão acumuladas sobre dez *folds*. Linhas representam as classes verdadeiras e colunas as classes preditas pelos modelos. A classe 0 representa compras atrasadas.

Os resultados do experimento são apresentados na Figura 4 e na Tabela 1. A Figura 4 traz as matrizes de confusão acumuladas entre *folds* para cada algoritmo avaliado. As linhas da matriz representam as classes verdeiras e as colunas são as classes preditas. A primeira classe (classe 0) indica que a conta sofreu atraso, enquanto a segunda classe (classe 1) indica que ela foi paga no prazo. Todos os modelos exceto o GNB alcançaram performance similar com valores de acurácia, precisão, revocação e F1-*score* próximos a 0.80. Os métodos com maiores valores médios das métricas foram o XGB e o AB.

O fato do classificador linear ter alcançado performance similar aos demais métodos é um indício de que o problema é linearmente separável dado o conjunto de características. Este resultado é compatível com o que é encontrado na literatura (Lessmann et al., 2015; Louzada et al., 2016; Dastile et al., 2020). Trabalhos anteriores realizados usando dados de bancos e financeiras apontam que classificadores lineares tipicamente são capazes de alcançar boa performance nas tarefas de análise de risco de crédito e pontuação de crédito.

4.2 Performance de Ensembles

Ensembles são conjuntos de classificadores que, em geral, fazem predições independentes para uma amostra de entrada e cujas predições são integradas usando algum tipo de algoritmo de sumarização. *Ensembles* possuem garantias teóricas de que sua performance será superior à dos classificadores de base dado que algumas condições de operação sejam satisfeitas (Bishop, 2006). Resultados empíricos comprovam a superioridade destes modelos na tarefa de análise de risco de crédito (Abellán and Castellano, 2017).

Neste experimento comparamos a performance de três ensembles construídos usando os classificadores listados na seção anterior. A sumarização foi realizada pelo número de votos. A classe mais votada pelos classificadores de base foi retornada. Os classificadores foram treinados independentemente e não foram usadas técnicas como *bagging* ou *boosting* para treinar o *ensemble*.

Tabela 2. Performance de *ensembles*.

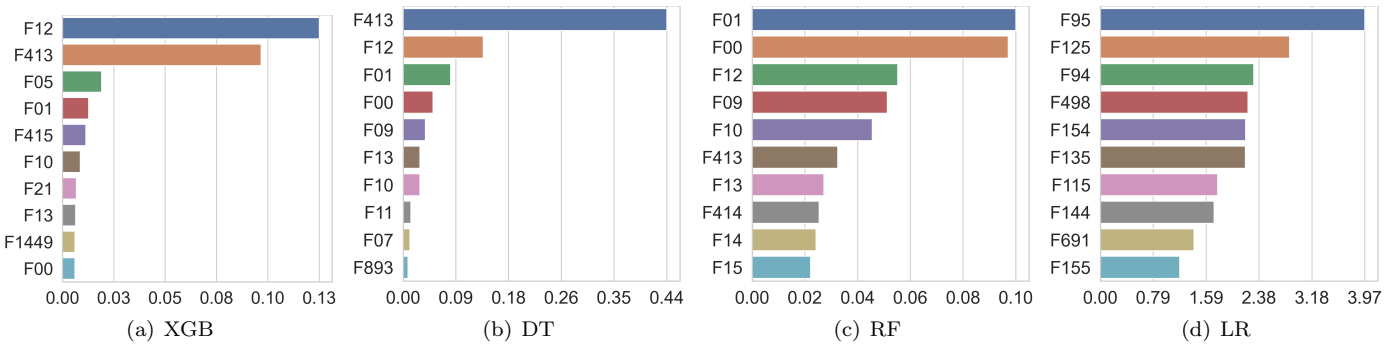
Modelo	Acurácia	Precisão	Revocação	F1-*Score*
E1	**0.81**	0.80	**0.80**	**0.80**
E2	**0.81**	**0.81**	**0.80**	**0.80**
E3	0.80	0.80	**0.80**	**0.80**

O *ensemble* E1 utilizou todos os modelos partindo do pressuposto que eles podem ter se especializado em partes diferentes da base de dados e que todas as contribuições são relevantes. O *ensemble* E2 utilizou os classificadores XGB, RF e AB dado que eles possuem a maior performance preditiva. Esta configuração assume que os métodos com baixa performance introduziriam mais ruído que contribuições relevantes. O *ensemble* E3 utilizou os modelos XGB, MLP e LR. Os primeiros modelos foram escolhidos por serem classificadores mais poderosos e que alcançam resultados estado-da-arte na literatura. O terceiro foi adicionado para evitar empates e ele foi selecionado por sua boa performance e característica complementar (hipótese de separabilidade linear).

Os resultados são apresentados na Tabela 2. Como pode ser observado, não houve ganho significativo de performance com as configurações de *ensembles* avaliadas. Este fato é um indício de que a performance dos classificadores únicos já estava próxima ao máximo que era possível para esta base de dados ou que todos aprenderam a identificar padrões espúrios similares de forma que ao uni-los não há redução de ruído.

4.3 Interpretação dos Modelos

Modelos de aprendizado de máquina podem encontrar padrões nos dados que levam à boa performance no conjunto de teste, mas que não necessariamente são regras desejáveis do ponto de vista humano. Geirhos et al. (2020) discutem o problema de redes neurais artificiais que alcançam boa performance nas bases de dados tipicamente usadas na literatura, mas não generalizam para o mundo real. Muitos dos argumentos valem para outros métodos de aprendizado de máquina. Hassani (2020) comenta o risco de propagação de preconceitos e desigualdade social pelo uso de aprendi-



(a) XGB (b) DT (c) RF (d) LR

Figura 2. Importâncias das características para os modelos baseados em árvores de decisão e contribuição das caraterísticas (veja o texto para mais detalhes) para os modelos lineares.

zado de máquina para a pontuação de crédito uma vez que estes modelos são construídos com dados históricos e não necessariamente justos.

Tendo em mente estas questões, esta seção investiga alguns dos modelos treinados para tentar identificar as regras aprendidas por eles. Para isto, foram consideradas a importância das características (Bishop, 2006) calculadas pelos modelos baseados em árvores de decisão e os pesos associados às características nos modelos lineares.

A Figura 4.3 traz gráficos representando as 10 características com maior importância para os modelos baseados em árvores de decisão e para o modelo linear a contribuição das características dadas pelos valores absolutos dos coeficientes multiplicados pelo desvio padrão das características. As descrições das características são apresentadas na Tabela 3.

Nos modelos XGB e DT, poucas características foram consideradas muito importantes para o modelo, enquanto nos modelos RF e LR, a importância e os valores dos coeficientes foram mais distribuídos. Isto significa que nos dois primeiros muitas amostras são classificadas usando poucas características de entrada, enquanto no segundo mais características são consideradas. Em geral, o uso de poucas características tende a indicar que o modelo encontrou regras rebuscadas e não naturais, mas ao mesmo tempo muito efetivas para resolver a tarefa. Assim, os comportamentos observados na RF e LR são mais desejáveis.

As características F12 e F413, as duas com maior importância no XGB e na DT, representam a data de realização da compra mais recente e a data do pagamento mais recente, respectivamente. É possível que os modelos tenham aprendido a comparar estes valores para avaliar se a pessoa pagou a compra anterior antes de fazer uma nova ou se houve atraso no pagamento. Outra possibilidade é que as datas indiquem a frequência ou intervalos de tempo entre compras e pagamentos.

As duas características com maior importância para a RF foram a F01 que indica a porcentagem de desconto no produto dado pelo vendedor e a F00 que armazena o saldo do cliente após a compra, i.e., qual seria o total da dívida a ser paga para a empresa se a compra for concretizada. Vendedores tendem a dar descontos para bons pagadores ou para convencer clientes a realizarem uma compra. Por outro lado, o saldo pós-compra é o tamanho da dívida do cliente e este valor pode indicar a situação financeira do cliente.

No caso da LR, os valores dos coeficientes estão bem distribuídos entre diversas características. As três características com maiores influências na saída foram são as F95 e F125 que indicam o valor de desconto dado ao cliente na oitava e décimas compras mais recentes. Já a F94, a terceira característica da lista representa o saldo final após a oitava compra mais recente. Como discutido no caso anterior, o desconto é um artifício usado por vendedores para aumentar a probabilidade de concretizar a venda e tipicamente são usados com bons clientes. O saldo, por outro lado, é um indicativo da situação financeira da pessoa com relação à empresa.

As análises realizadas nas duas seções permitem concluir que o modelo LR seria o melhor para implantação. Modelos lineares são menos suscetíveis à *overfitting* e permitem analisar os critérios usados na decisão. A principal limitação destes modelos é sua incapacidade de resolver problemas de classificação que não sejam linearmente separáveis. Como a performance do modelo linear é equivalente à dos demais, este não parece ser o caso na tarefa deste trabalho.

5. CONCLUSÃO

Este trabalho propôs e avaliou um sistema baseado em aprendizado de máquina para análise de risco de crédito em empresas gerenciadoras de vendedores porta a porta. Foi construída uma base de dados com dados demográficos de clientes, além de seu histórico de interações com a empresa como compras, pagamentos e agendamentos. Experimentos mostraram que alguns modelos de aprendizado de máquina são capazes de classificar compras para predizer se elas serão pagas no prazo acordado ou não com uma acurácia superior à 80%. O modelo com melhor performance preditiva foi o XGB com valores médios de acurácia, precisão, revocação e f1-*score* de 0.81, 0.80, 0.80 e 0.80, respectivamente. Vale mencionar, contudo, que a maioria dos outros modelos alcançou performance similar e que o classificador linear LR foi capaz de predizer se as compras seriam pagas no prazo ou não com a mesma performance dos modelos não lineares. Além disso, o uso de *ensembles* não trouxe benefícios significativos na base de dados utilizada.

Foram realizadas análise qualitativa das importâncias atribuídas às características pelos modelos baseados em árvore

Tabela 3. Características da Figura 4.3.

Nome	Descrição
F00	Dívida total do cliente se a compra for concretizada
F01	Valor de desconto na compra dado pelo vendedor
F05	Se a forma de pagamento não foi informada
F07	Se a foto do cliente foi registrada no sistema
F09	O gênero do cliente
F10	Se o CPF do cliente foi registrado no sistema
F11	Se o cliente foi cadastrado pelo vendedor porta-em-porta ou pela empresa gestora
F12	Data de realização da compra anterior
F13	Valor da compra anterior
F14	Saldo devedor após a compra anterior
F15	Desconto na compra anterior
F21	Se a forma de pagamento na compra anterior não foi registrada
F94	Saldo final da 8° venda mais recente
F95	Desconto dado à 8° venda mais recente
F115	Desconto dado à 10° venda mais recente
F125	Desconto dado à 11° venda mais recente
F135	Desconto dado à 12° venda mais recente
F144	Saldo final da 13° venda mais recente
F154	Saldo final da 14° venda mais recente
F155	Desconto dado à 14° venda mais recente
F413	Valor do pagamento mais recente
F414	Saldo devedor após o pagamento mais recente
F415	Se o pagamento mais recente foi um abatimento da dívida pela empresa gestora
F498	Saldo final do 7° do recebimento mais recente
F691	Correção de saldo 23° do recebimento mais recente
F893	Se o agendamento mais recente foi feito pelo aplicativo do vendedor porta em porta
F1449	Se o 46° agendamento anterior foi remarcado

de decisão e dos coeficientes do modelo linear para tentar identificar as regras de decisão aprendidas. A análise levou à conclusão de que em geral são consideradas características que fazem sentido para o problema. Contudo, o fato de que o XGB e a DT atribuíram muita importância à poucas características pode indicar o uso de regras não intuitivas do ponto de vista humano. Este tipo de regra tipicamente representa padrões espúrios na base de dados. Considerando os resultados de ambos os experimentos, o classificador LR é a melhor opção para implantação uma vez que ele é capaz de resolver o problema sendo menos suscetível à *overfitting* e por ser interpretável.

Embora sejam muito poderosos, hoje é difícil compreender precisamente as regras aprendidas por métodos de aprendizado de máquina. Em tarefas críticas como a análise de risco de crédito, esta característica pode levar à prejuízos para o credor e à julgamentos tendenciosos considerando o perfil do cliente. Assim, em trabalhos futuros será construído um sistema especialista de forma a obter um sistema cujas decisões possam ser justificadas. A performance preditiva do sistema especialista será avaliada usando a mesma base de dados para verificar se ele é verossímil considerando dados históricos.

REFERÊNCIAS

Abellán, J. and Castellano, J.G. (2017). A comparative study on base classifiers in ensemble methods for credit scoring. *Expert Systems with Applications*, 73, 1–10.

Banco Central do Brasil (2021). Sistema gerenciador de séries temporais - consultar séries. URL `https://www3.bcb.gov.br/sgspub/`.

Ben-David, A. and Frank, E. (2009). Accuracy of machine learning models versus "hand crafted" expert systems–a credit scoring case study. *Expert Systems with Applications*, 36(3), 5264–5271.

Bishop, C.M. (2006). *Pattern recognition and machine learning*. springer.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., et al. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4).

Dastile, X., Celik, T., and Potsane, M. (2020). Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing*, 91, 106263.

Federal Reserve System (2021). Consumer credit - g.19. URL `https://www.federalreserve.gov/releases/g19/current/`.

Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F.A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11), 665–673.

Hand, D.J. and Henley, W.E. (1997). Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 160(3), 523–541.

Hassani, B.K. (2020). Societal bias reinforcement through machine learning: a credit scoring perspective. *AI and Ethics*, 1–9.

He, H., Zhang, W., and Zhang, S. (2018). A novel ensemble method for credit scoring: Adaption of different imbalance ratios. *Expert Systems with Applications*, 98, 105–117.

Lessmann, S., Baesens, B., Seow, H.V., and Thomas, L.C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1), 124–136.

Louzada, F., Ara, A., and Fernandes, G.B. (2016). Classification methods applied to credit scoring: Systematic review and overall comparison. *Surveys in Operations Research and Management Science*, 21(2), 117–134.

Luo, C., Wu, D., and Wu, D. (2017). A deep learning approach for credit scoring using credit default swaps. *Engineering Applications of Artificial Intelligence*, 65, 465–470.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12, 2825–2830.

Pławiak, P., Abdar, M., Pławiak, J., Makarenkov, V., and Acharya, U.R. (2020). Dghnl: A new deep genetic hierarchical network of learners for prediction of credit scoring. *Information Sciences*, 516, 401–418.

Statista (2021). Value of consumer credit outstanding in the united states from 2000 to 2019. URL `https://www.statista.com/statistics/188170/consumer-credit-liabilities-of-us-households-since-1990/`.

Xia, Y., Liu, C., Li, Y., and Liu, N. (2017). A boosted decision tree approach using bayesian hyper-parameter optimization for credit scoring. *Expert Systems with Applications*, 78, 225–241.