II Workshop de Matemática, Estatística e Computação Aplicadas à Indústria

Técnicas de interpretabilidade para aprendizado de máquina: um estudo abordando avaliação de crédito

Daniel de Oliveira Caires[1]
Cláudio Fabiano Motta Toledo[2]
Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo

1 Introdução

Atualmente a inteligência artificial está sendo constantemente utilizada no dia a dia de pessoas e corporações, seja para prover informações relevantes, para apoiar no processo de decisão, apresentar sugestões, ou mesmo para diretamente realizar decisões importantes [2]. O aprendizado de máquina é, na maioria das vezes, o pilar para essas aplicações. Apesar das técnicas de aprendizado de máquina apresentarem altos níveis de acurácia e poderem ser aplicadas em vários campos da ciência, ainda existe certa dificuldade em confiar inteiramente nessas decisões sem o suporte em evidencias [1]. A busca por técnicas que permitam interpretar essas decisões podem trazer diversos benefícios. Além de ajudar ter maior confiança nas decisões apontadas, elas também podem evidenciar quais características estão tendo maior importância para determinado resultado, fazendo com que varias melhorias possam ser aplicadas tanto na fase de coleta de dados, evitando possíveis vieses, quanto no ajuste e treinamento do modelo. Este trabalho visa aplicar técnicas de interpretabilidade em um modelo preditivo de classificação, criado com algoritmo de aprendizado de máquina, para uma base de dados relacionada a concessão de crédito. Desta forma, faz parte do escopo do projeto o estudo de técnicas de aprendizado de máquina utilizadas nesses setores, as explorando no contexto da interpretabilidade.

2 Interpretabilidade em *Credit Score*

Para este trabalho, serão reproduzidas as etapas necessárias para a obtenção de um score de crédito como ilustrado na Figura 1. A partir de dados históricos, são aplicados os pré-tratamentos necessários (Etapa 1) para posterior treinamento com as técnicas de aprendizado de máquina (Etapa 2). Os resultados serão avaliados de acordo com métricas de performance adotadas. O

[1]dielcaires@usp.br
[2]claudio@icmc.usp.br

problema central que esta dissertação aborda esta na interpretação do impacto das variáveis para os resultados do modelo. Assim, as técnicas de interpretabilidade serão aplicadas (Etapa 3). Ao final do trabalho, deseja-se apresentar um entendimento claro de quão determinada variável ou conjunto de variáveis influenciaram para que determinado resultado fosse obtido. Tal entendimento será baseado em métricas providas pelas técnicas de interpretabilidade utilizadas. Uma vez o modelo finalizado, avaliado e interpretado, ele pode ser implantado e disponibilizado para o cliente (Etapa 4).

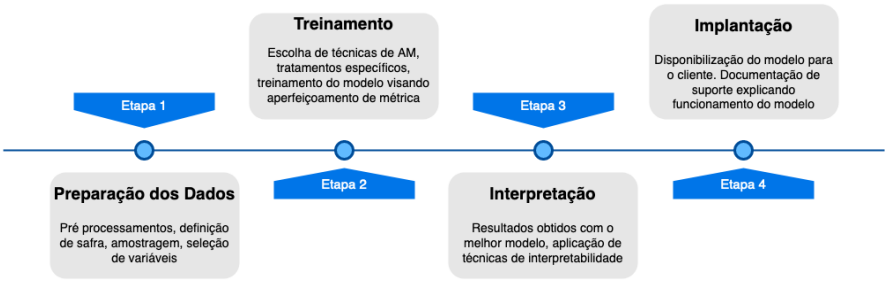

Figura 1: Ilustração mostrando as etapas do *Credit* Score

Para este trabalho o treinamento do modelo foi feito com o algoritmo de aprendizado de máquina chamado *LightGBM*. Ele possui uma biblioteca na linguagem *Python*. Sua escolha se deve a eficiência de tempo de processamento e grande acurácia na predição [3]. Esse modelo é baseado no *Gradient Boosting*, técnica que utiliza *Ensembles*, que são classificadores que fazem uma combinação de resultados de preditores fracos, como arvores de decisão, com o objetivo de produzir um melhor modelo preditivo. A técnica de interpretabilidade escolhida, o *SHAP* também possui implementação em uma biblioteca em *Python* [4]. No problema abordado neste trabalho, a base de dados é estruturada, tratando-se de uma amostra aleatória com informações financeiras de consumidores brasileiro, totalizando 750 mil consumidores, datando entre maio de 2020 a junho de 2021. Dentre as informações presentes, estão os dados típicos de um bureau de crédito, que são descritos na Tabela 1. A variável resposta do problema é o chamado conceito de mercado - uma definição binária que define um consumidor como um bom ou mau tomador de crédito diante do mercado, de acordo com seu comportamento em um período histórico analisado.

3 Resultados

O modelo desenvolvido foi avaliado usando a técnica do *SHAP*. Ele permite avaliar globalmente e localmente o modelo através de diferentes visualizações. A métrica de importância medida por essa técnica, chamada de *SHAP Value*, pode ter um valor positivo ou negativo. Com o módulo dessa medida, foi obtido o gráfico de barras da Figura 2. A variável *Var_Qtde_Restr_1* foi a que teve maior importância. Faz sentido que ela seja bastante relevante ao se considerar o risco associado a esse consumidor, uma vezes que um consumidor com muitas dividas tende a ser mais arriscado. A Figura 3 mostra um gráfico que analisa a variável *Var_Qtde_Restr_1* individualmente, comparando os valores dessa variável no eixo x com os valores *SHAP* no eixo y. Essa variável contínua e faz parte do grupo de variáveis de restrição. É possível observar que existe uma região

Tabela 1: Descrição da Base de Dados

Bloco variáveis	Tipo	Valores
Valor Pagamento	Contínua - números reais (valor em R$)	de 0 a +inf
Pontualidade	Contínua - percentual (% do valor)	de 0 a 1
Dias de atraso	Contínua - inteiro (dias)	de 0 a +inf
Tempo de contratação	Contínua - inteiro (dias)	de 0 a +inf
Parcelas Pagas	Contínua - inteiro (quantidade parcelas)	de 0 a +inf
Parcelas Atrasadas	Contínua - inteiro (quantidade parcelas)	de 0 a +inf
Quantidade de Negativações	Contínua - inteiro (quantidade negativ.)	de 0 a +inf
Valor das Negativações	Contínua - números reais (valor em R$)	de 0 a +inf
Quantidade de Consultas	Contínua - inteiro (quantidade consult.)	de 0 a +inf
Indicador de Pré Aprovado	Binária (0 - Não e 1 - Aprovado)	0 e 1
Variável Resposta: Conceito	Binária (0 - Bons e 1 - Maus)	0 (69.3%) e 1 (30.7%)

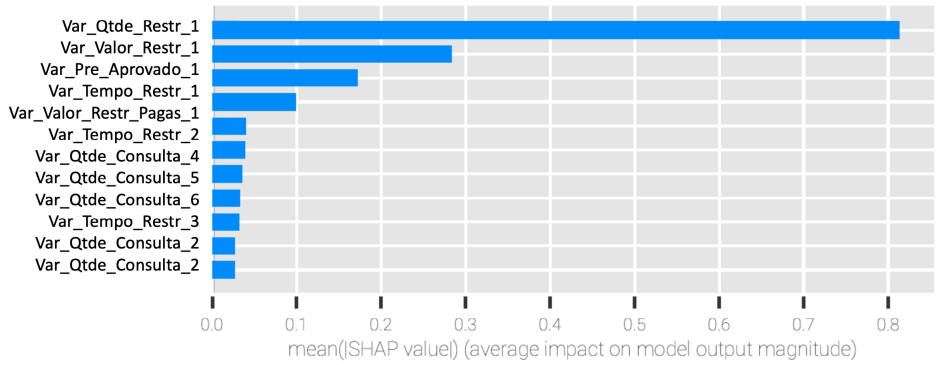

Figura 2: Importância por Shap na base de crédito.

(circulada com a cor verde) onde os valores são 0 pois estão colados a linha do eixo y, e estão associados a valores *SHAP* maiores, indicando impacto positivo. Quando são maiores, em geral se concentram em uma região (circulada com a cor vermelha) onde os valores *SHAP* são menores que zero (entre -0.5 e -2) indicando um impacto negativo bem considerável.

Também é possível fazer analises locais, estudando casos isolados para entender quais fatores levaram a determinada classificação. A Figura 4 mostra um gráfico *Waterfall* para um caso que foi estimado como um bom pagador. Para este caso, a variável *Var_Qtde_Restr_1* teve o maior impacto positivo, correspondente a +0.41 de *SHAP value*, uma vez que ela teve valor 0. A segunda variável de maior impacto, foi a *Var_Valor_Restr_1*, que também teve valor 0, adicionando +0.31 pontos.

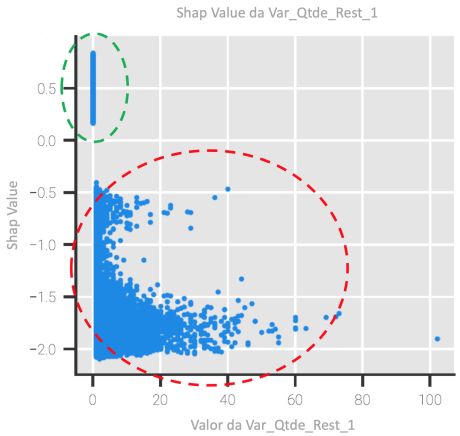

Figura 3: Gráfico de dispersão de valor *SHAP* para variável de restrição.

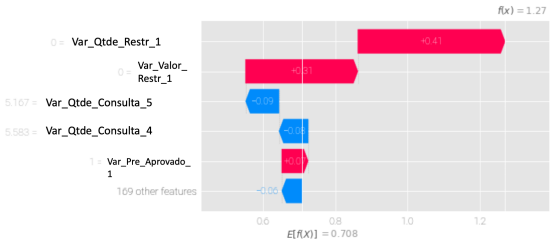

Figura 4: Gráfico *Waterfall* analisando localmente caso bom.

4 Conclusões

Fica claro que a interpretabilidade pode servir com um grande catalisador para a adoção de aprendizado de máquina nos mais diversos contextos. Para o modelo de crédito, foi possível mostrar que fatores como negativações anteriores, pre aprovações e quantidade de consultas a crédito foram relevantes para o modelo. Um gestor de um produto de crédito, com posse dessas informações, poderia ajustar melhor suas políticas internas, ou mesmo solicitar novos modelos, de acordo com suas necessidades, a partir dos conhecimentos obtidos com essa análise. Ele poderia querer trabalhar com um publico um pouco mais arriscado, e considerar valores baixos de negativação, ou negativações muito antigas, como fatores que não impedissem a concessão, por exemplo. Foi possível no entanto observar certas limitações. Mesmo com métricas e visualizações muito úteis ainda não foi possível obter uma formula direta que remeta a probabilidade obtida com a predição do modelo, como acontece em técnicas mais simples e interpretáveis.

Referências

[1] Chen, J. H.; ASCH, S. M. *Machine learning and prediction in medicine—beyond the peak of inflated expectations. The New England journal of medicine, NIH Public Access*, 2017.

[2] Grgic-Hlaca, N.; Zafar, M. B.; Gummadi, K. P.; Weller, A. *The case for process fairness in learning: Feature selection for fair decision making. NIPS Symposium on Machine Learning and the Law.*, 2016.

[3] Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. *LightGBM: A highly efficient gradient boosting decision tree.*, 2017.

[4] Lundberg, S. M. *SHAP (SHapley Additive exPlanations)*. 2017. Acessado em 22/06/2022. Disponível em: ¡https://github.com/slundberg/shap¿.