UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Interpretação de modelos complexos de aprendizado de máquina

Davi Keglevich Neiva

Dissertação de Mestrado do Programa de Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria (MECAI)

ICMC USP
SÃO CARLOS

Davi Keglevich Neiva

Interpretação de modelos complexos de aprendizado de máquina

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Mestre – Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria. *VERSÃO REVISADA*

Área de Concentração: Matemática, Estatística e Computação

Orientador: Prof. Dr. Paulino Ribeiro Villas Boas

USP – São Carlos
Dezembro de 2023

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

K26i

Keglevich Neiva, Davi
Interpretação de modelos complexos de aprendizado de máquina / Davi Keglevich Neiva; orientador Paulino Ribeiro Villas Boas. -- São Carlos, 2023.
73 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2023.

1. Aprendizado de Máquina. 2. Valores SHAP. 3. Interpretabilidade de modelos. I. Ribeiro Villas Boas, Paulino , orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Davi Keglevich Neiva

Complex machine learning models interpretation

Dissertation submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP – in accordance with the requirements of the Professional Master's Program in Mathematics Statistics and Computing Applied to Industry, for the degree of Master in Science. *FINAL VERSION*

Concentration Area: Mathematics, Statistics and Computing

Advisor: Prof. Dr. Paulino Ribeiro Villas Boas

USP – São Carlos
December 2023

Dedico este a trabalho ao meu pai e à minha irmã, que sempre me deram apoio incondicional, além de terem sido grandes exemplos a serem seguidos. Vocês estarão sempre comigo.

AGRADECIMENTOS

Agradeço a todos os professores que participaram na minha formação como profissional e ser humano.

Aos colegas de trabalho, por proporcionarem um ambiente de crescimento e aprendizado. Aos amigos, por todos os momentos felizes.

Ao meu pai e à minha irmã, por terem sido pessoas incríveis e minhas grandes referências. À minha esposa e companheira de vida, que, além de fazer a vida ser mais leve, sempre me apoia e faz com que os desafios sejam mais fáceis de serem superados.

“Para ser ignorante, uma pessoa precisa estudar muito;
senão, é só inocente. ”
(Wilton Bussab)

RESUMO

NEIVA, D. K. **Interpretação de modelos complexos de aprendizado de máquina**. 2023. 73 p. Dissertação (Mestrado – Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023.

Algoritmos de aprendizado de máquina são amplamente utilizados em diversos setores da sociedade e desempenham um papel significativo na tomada de decisões em vários contextos. Esses algoritmos são capazes de criar modelos cada vez mais sofisticados, que conseguem capturar relações complexas nos dados para alcançar resultados mais acurados. No entanto, à medida que esses modelos desenvolvem relações mais complexas, a compreensão de seu funcionamento também se torna mais desafiadora. Esses modelos de aprendizado de máquina frequentemente incorporam centenas, ou até mesmo milhares, de variáveis. Neste trabalho, apresentamos alguns algoritmos de aprendizado de máquina, abordamos sua complexidade e discutimos a importância de compreender o funcionamento desses modelos complexos. Além disso, exploramos a metodologia *SHAP* para interpretar modelos de *boosting* (classificação e regressão) em 3 estudos de caso distintos: identificação dos perfis mais propensos a alcançarem uma nota mínima no ENEM - Exame Nacional do Ensino Médio; desenvolvimento de um *score* de risco de crédito de uma cooperativa de empresas e avaliação da concentração de carbono em amostras de solo de diferentes biomas brasileiros a partir de dados de espectroscopia. Com a utilização da metodologia *SHAP* foi possível trazer informações complementares às do modelo em cada um desses casos, revelando padrões de características socio econômicas dos candidatos do ENEM, características das empresas que o modelo aprendeu no desenvolvimento do *score* de crédito e informações relevantes sobre a composição dos solos. A interpretação dos modelos não apenas aprimora a análise dos conjuntos de dados, mas também possibilita a identificação de vieses amostrais, a avaliação do aprendizado obtido durante a construção dos modelos e, até mesmo, a revelação de informações que podem não ser prontamente discerníveis nos dados.

Palavras-chave: Interpretabilidade de modelos, Aprendizado de máquina, Modelos complexos, Valores *SHAP*.

ABSTRACT

NEIVA, D. K. **Complex machine learning models interpretation**. 2023. 73 p. Dissertação (Mestrado – Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023.

Machine learning algorithms are widely used in various sectors of society and play a significant role in decision-making in various contexts. These algorithms are capable of creating increasingly sophisticated models that can capture complex relationships in data to achieve more accurate results. However, as these models develop more complex relationships, understanding how they work also becomes more challenging. Machine learning models often incorporate hundreds, or even thousands, of variables. In this work, we present some machine learning algorithms, discuss their complexity, and emphasize the importance of understanding the functioning of these complex models. Furthermore, we explored the *SHAP* methodology to interpret boosting models (classification and regression) in three distinct case studies: identifying profiles most likely to achieve a minimum score on the ENEM - National High School Exam; developing a credit risk score for a cooperative of companies, and evaluating carbon concentration in soil samples from different Brazilian biomes using spectroscopy data. With the use of the *SHAP* methodology, it was possible to provide additional information to the model in each of these cases, revealing patterns of socioeconomic characteristics of ENEM candidates, characteristics of the companies that the model learned in the development of the credit score, and relevant information about soil composition. The interpretation of the models not only enhances the analysis of the datasets but also allows for the identification of sample biases, evaluation of the learning acquired during model construction, and even the revelation of information that may not be readily discernible in the data.

Keywords: Model interpretability, Machine learning, Complex models, *SHAP* values.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

Nos últimos anos, o avanço tecnológico tanto de hardware quanto de software possibilitou a aplicação e desenvolvimento de técnicas mais sofisticadas de aprendizado de máquina. Essas técnicas são utilizadas em várias aplicações práticas, como carros autônomos, filtros de e-mail e sistemas de previsão usados por departamentos de polícia (DOSHI-VELEZ; KIM, 2017). De acordo com Coglianese e Lehr (2016), algoritmos de aprendizado de máquina estão revolucionando diversos setores da economia, incluindo motores de busca, processamento de imagens médicas e muitas outras aplicações.

Segundo Brynjolfsson e Mitchell (2017) nós estamos vivendo em um momento de grande transformação devido à evolução na área de aprendizado de máquina, que está possibilitando uma aceleração da automação em diversas áreas. O autor argumenta que alguns dos impactos desse processo são: substituição do trabalho humano em algumas atividades por sistemas desenvolvidos através de técnicas de aprendizado de máquina; alteração nos preços de produtos e serviços, devido à redução dos custos através dessas novas técnicas; aumento da demanda de trabalho em atividades que exijam uma maior qualificação (atividades que não foram automatizadas) e a alteração dos planos de negócio de diversas empresas. Além de alterar toda uma estrutura de produção de bens e serviços, o que fica claro é que essas tecnologias estão cada vez mais presentes na sociedade e são importantes no processo de decisão tanto para empresas quanto para pessoas.

Em muitas atividades o desempenho de sistemas desenvolvidos através de aprendizado de máquina supera o de humanos, principalmente quando a atividade envolve analisar grandes volumes de dados. Nesse sentido, muitas vezes esses sistemas são utilizados para guiar os humanos em processos de compreensão sobre algum determinado tema e de decisão sobre alguma ação (DOSHI-VELEZ; KIM, 2017). Em razão disso, a falta de compreensão e clareza sobre o funcionamento desses sistemas passa a ser um problema, principalmente em áreas mais críticas como medicina, justiça criminal, sistemas financeiros, entre outros (LIPTON, 2018).

Uma área mais específica em que esse tema é muito relevante é o mercado de crédito. Esse mercado é de grande importância para a economia, ele possibilita a aceleração do consumo e da produção, o que é fundamental para promover o crescimento econômico. Ele tem como base a necessidade que as pessoas físicas e jurídicas têm, no curto e médio prazo, de obterem um maior poder de compra para adquirir bens e serviços (OLIVEIRA, 2018). No entanto, é necessária uma boa gestão de riscos para que a concessão de crédito não seja feita de uma forma insustentável e que a inadimplência se mantenha em níveis razoáveis. É nesse contexto que são utilizados modelos que auxiliam a tomada de decisão, através de um indicador (chamado de *score*) que apresenta a propensão de cada indivíduo cumprir seus compromissos pecuniários. A partir desse *score*, as empresas elaboram uma política que vai definir as pessoas que vão ter acesso ao crédito. Diante disso, grandes empresas que atuam nesse mercado têm investido cada vez mais em soluções que envolvem modelos de aprendizado de máquina, visando alcançar uma maior assertividade na tomada de decisões. Porém, devido a sua natureza complexa, há ainda uma certa desconfiança no ambiente financeiro (FORTI, 2018).

Harari (2018), no seu livro *21 Lições para o século 21* apresentou, de forma bastante didática, a importância desse assunto e quais podem ser algumas das implicações da utilização desses mecanismos na tomada de decisões:

> (...) Mesmo que a democracia consiga se adaptar e sobreviver, as pessoas podem tornar-se as vítimas de novos tipos de opressão e discriminação. Já hoje em dia, cada vez mais bancos, corporações e instituições estão usando algoritmos para analisar dados e tomar decisões a nosso respeito. Quando você pede um empréstimo a seu banco, é provável que seu pedido seja processado por um algoritmo e não por um humano. O algoritmo analisa grande quantidade de dados sobre você e estatísticas sobre milhões de outras pessoas, e decide se você é confiável o bastante para receber um empréstimo. Frequentemente, o algoritmo faz o trabalho melhor do que faria um gerente. Mas o problema é que se o algoritmo discriminar injustamente certas pessoas, será difícil saber. Se o banco se recusar a lhe dar um empréstimo e você perguntar por quê, o banco responderá: "O algoritmo disse que não". Você pergunta: "Por que o algoritmo disse não? O que há de errado comigo?", e o banco responde: "Não sabemos. Nenhum humano entende esse algoritmo, porque é baseado num aprendizado de máquina avançado. Mas confiamos em nosso algoritmo, por isso não lhe daremos um empréstimo".
>
> Quando se discriminam grupos inteiros, como mulheres ou negros, esses grupos podem se organizar e protestar contra a discriminação coletiva. Mas agora um algoritmo seria capaz de discriminar você individualmente, sem que você saiba por quê. Talvez o algoritmo tenha encontrado alguma coisa da qual não gostou em seu DNA, em sua história pessoal ou em sua conta no *Facebook*. O algoritmo teria discriminado você não porque é mulher ou negro — mas porque você é você. Há algo específico em você de que o algoritmo não gosta. Você não sabe o que é, e mesmo se soubesse não poderia organizar um protesto com outras pessoas, porque não há outras pessoas que sejam alvo do mesmo preconceito. É só você.

> Em vez de só discriminação coletiva, no século XXI talvez deparemos com um crescente problema de discriminação individual.
>
> Nos níveis mais altos da autoridade provavelmente ainda teremos figurantes humanos, que nos darão a ilusão de que os algoritmos são apenas conselheiros, e que a autoridade final ainda está em mãos humanas. Não vamos nomear uma IA chanceler da Alemanha ou CEO do *Google*. No entanto, as decisões tomadas pelo chanceler da Alemanha ou pelo CEO do *Google* serão formuladas pela IA. O chanceler ainda poderia escolher entre várias opções diferentes, mas todas seriam resultado da análise feita por Big Data, e refletirão mais como a IA vê o mundo do que como os humanos o veem. [1](HARARI, 2018)

Entender como essas ferramentas de aprendizado de máquina são construídas e evidenciar como elas estão funcionando é fundamental para evitar um cenário como esse descrito por Harari (2018). Muitas técnicas de aprendizado de máquina se baseiam em modelos que aplicam abordagens não paramétricas e não lineares, focadas principalmente em aumentar o poder preditivo e que resultam em sistemas mais precisos e menos interpretáveis (MOLNAR; CASALICCHIO; BISCHL, 2020). Unceta, Nin e Pujol (2018) também apontam que esses tipos de modelos são capazes de capturar relações mais complexas entre as variáveis preditoras e, por isso, entregar predições consideravelmente mais precisas. No entanto, os autores ressaltam que o uso dessas abordagens no mercado de crédito está sujeito a auditoria e fiscalização no sentido de averiguar quais são as regras de decisão e muitas vezes são exigidas especificações técnicas de quais são as variáveis mais importantes e como elas se relacionam. Por essa razão, essa exigência de interpretabilidade inviabiliza a utilização dessas técnicas no mercado de crédito, fazendo com que grandes empresas continuem utilizando técnicas mais clássicas, como a de regressão logística, que não têm o mesmo potencial de assertividade, mas que são mais simples de serem interpretadas.

De acordo com Lundberg e Lee (2017) em muitos contextos a interpretação de um modelo pode se tornar um fator tão importante quanto a precisão de suas predições. Esses autores também afirmam que a interpretabilidade do modelo, além de permitir um melhor entendimento e gerar uma maior confiança para decisões embasadas em suas predições, pode ser crucial para identificar regras de decisões equivocadas ou até mesmo formas de aprimorar sua performance. Um modelo mais claro possibilita uma melhor compreensão do próprio funcionamento do negócio.

Diante dessa necessidade de entender a construção desses modelos e explicar as particularidades de um modelo específico, muitas pesquisas foram endereçadas a esse desafio. De acordo com Molnar, Casalicchio e Bischl (2020) a área que busca estudar o aprendizado de máquina interpretável (frequentemente chamada de *IML*, do inglês *Interpretable Machine Learning*) realmente alcançou notoriedade a partir de 2015, quando a quantidade de buscas por termos relacionados no google e a quantidade de artigos publicados sobre o tema aumentou

[1] Harari, Y. N. (2018). 21 lições para o século 21. Editora Companhia das Letras. Página 70.

substancialmente. Esses autores apontam também que muitas pesquisas e metodologias criadas com foco em interpretabilidade de machine learning são recentes e que ainda há muito o que explorar, mas hoje já existem softwares livres dedicados a *IML* e que há um esforço em aplicar essas técnicas na indústria, inclusive com *startups* direcionadas especificamente a esse tipo de desafio e grandes empresas produzindo softwares próprios.

Mesmo com um grande número de pesquisadores concentrados em desenvolver métodos que permitam interpretabilidade de modelos complexos, ainda não há uma solução que é amplamente aceita e difundida. Molnar, Casalicchio e Bischl (2020) afirmam que hoje há uma maior compreensão de quais são os pontos fracos e limitações dos métodos existentes para interpretabilidade de aprendizado de máquina e que cada vez mais surgem novas técnicas que apresentam níveis de complexidade ainda maiores, o que indica que esse tema tende a ser mais discutido e estudado.

Neste trabalho será abordada uma metodologia para interpretar o funcionamento de modelos desenvolvidos com uma técnica de aprendizado de máquina. Além disso, será analisada a implementação de modelos de aprendizado e suas respectivas interpretações em 3 cenários distintos: no conjunto de dados socieconômicos de participantes do ENEM; na construção de um modelo de concessão de crédito através de dados financeiros reais e na predição da concentração de carbono em amostras de solo reais a partir de dados de espectroscopia. Com isso, o objetivo do trabalho é identificar a importância e utilidade da interpretação de modelos em diferentes contextos de modelagem. Seja o propósito do modelo inferencial ou preditivo, a busca de uma boa interpretação pode servir como auxílio para um melhor uso do modelo.

CAPÍTULO

2

APRENDIZADO DE MÁQUINA

Antes de compreender quais são as metodologias que estão sendo desenvolvidas para interpretar modelos de aprendizado de máquina, é importante contextualizar como esses modelos são criados.

2.1 O que é aprendizado de máquina?

De acordo com Athey (2018) o termo 'aprendizado de máquina' tem sido amplamente utilizado e não apresenta uma definição única, ele pode ser relacionado a áreas de estudo de ciências da computação, mas também a pesquisas que envolvem estatística, engenharia, ciências sociais, entre outras. A autora apresenta uma das definições de aprendizado de máquina como uma área que desenvolve algoritmos que têm como objetivo criar modelos que possam realizar predições (através de uma regressão, classificação e/ou agrupamentos).

Naqa e Murphy (2015) apontam que os algoritmos de aprendizado de máquina buscam executar alguma tarefa através de um processo computacional que utilize conjuntos de dados e que não precisam ser especificamente programados para isso. Em outras palavras, esses algoritmos têm a capacidade de 'aprender' a partir dos dados que foram disponibilizados. Eles ainda apresentam que esses algoritmos visam emular a forma que seres humanos aprendem a realizar tarefas como a de identificação de padrões.

Os algoritmos de aprendizado de máquina podem ser divididos em dois grandes grupos, os supervisionados e os não supervisionados. O primeiro grupo é relacionado a conjuntos de dados que têm o seu respectivo valor de saída observado e os algoritmos desse grupo podem então ajustar os seus parâmetros para apresentar saídas próximas ao que é esperado. Tarefas de classificação, em que o algoritmo tem o objetivo de identificar quais são as características mais relacionadas a uma categoria específica são exemplos que se encaixam no contexto de aprendizagem supervisionada. Já o segundo grupo contém algoritmos que não necessitam de um

valor esperado e buscam identificar grupos que apresentam maior similaridade (ATHEY, 2018).

Nas próximas seções serão brevemente descritas algumas das técnicas de aprendizado de máquina mais utilizadas para poder dimensionar o grau de complexidade que esses modelos podem atingir.

2.2 Técnicas de modelagem

Coglianese e Lehr (2016) afirmam que as técnicas de aprendizado de máquina são não-paramétricas, o que significa que não requerem que o pesquisador defina uma função matemática para representar os dados. Em vez disso, elas usam as informações disponíveis para ajustar um modelo através de um algoritmo, permitindo que aprendam relações complexas. Isso resulta em modelos não lineares, que são capazes de aprender a partir de dados com alta complexidade.

2.2.1 Árvores de decisão

As árvores de decisão são uma técnica de aprendizado de máquina que se destacam pela sua relativa facilidade de compreensão. Ao contrário de outras técnicas, elas se baseiam em um número reduzido de relações entre os dados disponíveis. Segundo Podgorelec *et al.* (2002), as árvores de decisão são um método de aprendizado supervisionado que consiste em nós conectados a subárvores, relacionando os atributos dos dados para predizer o valor de saída com precisão. Durante o ajuste da árvore, são selecionados os atributos que melhor separam os dados, de forma a minimizar uma métrica de erro. Alguns termos comuns que descrevem um modelo de árvore são: profundidade, nós, folhas e ramos. A profundidade refere-se ao número de níveis em que os conjuntos de dados podem ser separados na árvore - quanto mais profunda a árvore, mais complexa ela é. Um nó é o resultado de uma separação apresentada pela decisão do nível acima da árvore. As folhas apontam para um nó terminal que não apresenta mais decisões e, portanto, representam grupos de dados com características em comum. Já um ramo é o conjunto de decisões subjacentes a um nó. As árvores de decisão são importantes porque servem de base para outros algoritmos de aprendizado de máquina, como será comentado a seguir.

2.2.2 Florestas aleatórias

Apesar de menos complexas, as árvores de decisão usualmente apresentam um menor poder preditivo se comparado com outras técnicas (IZBICKI; SANTOS, 2020). Diante disso, uma forma de alavancar a assertividade é utilizar mais de um modelo de forma conjunta através de uma técnica de *ensemble*, combinação de vários modelos preditivos com o objetivo de obter predições mais precisas e estáveis, como a de *bagging* (abreviação de *bootstrap aggregating*) onde são obtidos diversos modelos a partir de subconjuntos de dados para então agregar todas essas predições em uma única (GÉRON, 2017). Izbicki e Santos (2020) apontam que a técnica de florestas aleatórias melhoram a predição de árvores através do *ensemble* de diversas árvores,

resultando em modelos mais precisos, mas menos interpretáveis. Essas florestas aleatórias podem envolver milhares de árvores de decisão que foram ajustadas a partir de amostragens aleatórias do conjunto de dados e têm como predição a média das predições de todas essas árvores, o que evidencia a complexidade desses modelos (COGLIANESE; LEHR, 2016).

2.2.3 Impulsionamento de árvores de decisão

Uma outra técnica de aprendizado de máquina bastante utilizada é o de impulsionamento de árvores de decisão, mais conhecido como *boosting*. Essa técnica também se trata de uma forma de *ensemble*, reunindo diversos modelos de árvores de decisão em um único. No entanto, diferentemente das florestas aleatórias esses modelos são adicionados de forma sequencial, com uma árvore de decisão buscando corrigir os erros observados da anterior (GÉRON, 2017). Assim como as florestas aleatórias esses modelos podem envolver diversas árvores, o que normalmente melhora a sua precisão, mas aumenta consideravelmente a sua complexidade. De acordo com Izbicki e Santos (2020), há várias implementações desse tipo de modelo, como o *XGBoost*. Nesse trabalho será utilizado principalmente modelos de Impulsionamento de árvores de decisão, que são amplamente utilizados e apresentam performances consideradas como estado da arte (CHEN; GUESTRIN, 2016).

2.2.4 K-médias

Diferentemente das outras técnicas mencionadas de aprendizado de máquina, as K-médias são uma técnica de aprendizado não supervisionada e que busca agrupar o conjunto de dados em k-grupos a partir das similaridades que foram identificadas pelo algoritmo. Essas similaridades são calculadas a partir da distância de k-centros (chamados de centróides), que são atualizados de forma iterativa até que esses centros não se alterem e os grupos identificados continuem os mesmos entre duas iterações (PALMA, 2018). Ainda segundo Palma (2018), essa técnica é amplamente aplicada em mineração de dados, estatística, engenharia, aprendizado de máquina entre outras.

2.2.5 Redes neurais profundas

Redes neurais profundas são um tipo de aprendizado de máquina que envolve múltiplas camadas de redes neurais artificias que são interconectadas. As camadas que não estão na entrada ou na saída do modelo são chamadas de camadas ocultas e todas as camadas podem envolver diferentes quantidades de neurônios, a depender da arquitetura da rede em questão (IZBICKI; SANTOS, 2020). Essas redes buscam emular como um cérebro biológico funciona e identificar padrões em atividades complexas, como reconhecimento de fala, visão computacional e processamento de linguagem natural (LECUN; BENGIO; HINTON, 2015). Nesse trabalho não serão abordados modelos de redes neurais profundas, mas existem muitas metodologias

sendo desenvolvidas para compreender como são as características de um modelo desenvolvido com esse tipo de técnica e explicar suas respectivas predições.

São várias as formas de desenvolver modelos de aprendizado de máquina e cada vez mais os pesquisadores têm criado variações e novas técnicas que podem apresentar maior precisão nas tarefas que elas são destinadas. No entanto, é evidente que o que todas essas técnicas têm em comum é uma complexidade que, principalmente quando associadas a grandes conjuntos de dados, dificultam a compreensão de como cada tarefa específica é feita.

CAPÍTULO

3

INTERPRETAÇÃO DE APRENDIZADO DE MÁQUINA

3.1 O que é interpretabilidade de um modelo?

Não há uma definição técnica e formal do que é interpretabilidade no contexto de modelos de aprendizado de máquina. Alguns autores afirmam que interpretabilidade está relacionada a compreensão de como o modelo funciona, enquanto outros argumentam que o termo é ligado à capacidade de explicar as suas predições (LIPTON, 2018). O primeiro entendimento é associado a esclarecer o mecanismo que foi construído através do aprendizado de máquina, enquanto o segundo pode envolver soluções que explicam as predições feitas pelo modelo, mas não necessariamente elucidam como esse modelo funciona. Segundo (MOLNAR; CASALICCHIO; BISCHL, 2020) o conceito de interpretabilidade é subjetivo, não há uma maneira de mensurar o quanto que um modelo é interpretável ou não, mas já existem estudos para quantificar alguns dos aspectos desse modelo, como força de interação entre variáveis, sensitividade a perturbações dos dados de entrada, capacidade de obter predições em um conjunto de dados específico (simulabilidade), o quanto que a explicação se aproxima das predições (fidelidade) entre outras. O que fica evidente é que não só os modelos alcançaram um maior nível de complexidade, a discussão sobre como interpretá-los também.

3.2 Orientado a modelo ou agnóstico?

Algumas técnicas de modelagem têm a característica de gerar modelos que são mais fáceis de serem compreendidos, como os de regressão logística, regressão linear e árvores de decisão, porque a própria estrutura do modelo, através de seus coeficientes/regras de decisão, já permitem interpretar o seu funcionamento. Para essas técnicas a análise desses componentes do modelo já é uma abordagem para entender e apresentar o seu funcionamento. Porém, mesmo

essas metodologias quando aplicadas a grandes quantidades de variáveis preditoras podem gerar modelos muito mais complexos e de difícil interpretação (MOLNAR; CASALICCHIO; BISCHL, 2020). Molnar, Casalicchio e Bischl (2020) também comentam que nesses casos há técnicas que buscam reduzir complexidade, como o *LASSO* e imposição de restrições de monotonicidade das variáveis preditoras, ou propõem métricas que quantificam importância de variáveis, visando facilitar sua interpretação através da análise de seus componentes. Um outro exemplo de metodologia que é vinculada a uma técnica específica é o mapa de saliência, que visa trazer interpretação especificamente para redes neurais convolucionais (MOLNAR; CASALICCHIO; BISCHL, 2020). Essas abordagens, por considerarem as estruturas específicas dos modelos, são consideradas orientadas a modelo.

Diante desses modelos que apresentam relações complexas entre as variáveis preditoras, foram criadas metodologias que geram modelos substitutos com a finalidade de indicar o comportamento de um modelo complexo de uma forma mais simples. Os modelos substitutos examinam os modelos mais complexos como um sistema fechado com o objetivo de relacionar as informações de entrada com as predições desse sistema da forma mais precisa e simplificada possível (KIM; BOUKOUVALA, 2019). Esses autores também argumentam que uma vez que esse modelo substituto é ajustado é possível compreender o comportamento do modelo complexo através da análise de seus componentes. Molnar, Casalicchio e Bischl (2020) afirmam que muitos dos métodos que buscam possibilitar a interpretação de modelos complexos se baseiam em gerar modelos mais simplificados e que têm a característica de serem agnósticos (não levam em conta qual é a estrutura do modelo complexo, uma vez que ele é enxergado como um sistema fechado) e portanto são mais abrangentes. Um exemplo é o *LIME* (*Local interpretable model-agnostic explanations*), onde se busca trazer explicações individuais entre as predições do modelo complexo, apresentando o comportamento desse modelo especificamente para uma região das variáveis preditoras, sendo válido para indivíduos com características próximas.

3.3 Explicabilidade local e global

Muitas metodologias de interpretabilidade de modelos de aprendizado de máquina envolvem analisar as alterações das predições do modelo a partir de variações inseridas no conjunto de dados de entrada (MOLNAR; CASALICCHIO; BISCHL, 2020). É importante compreender, portanto, os conceitos de explicações locais e globais.

A explicabilidade local ocorre quando uma predição específica é explicada. É importante analisar o modelo localmente porque características dos dados que são importantes para o modelo como um todo podem não ser em casos menos gerais, e vice-versa (RIBEIRO; SINGH; GUESTRIN, 2016). Além do *LIME*, Molnar, Casalicchio e Bischl (2020) apontam as explicações contrafactuais como uma forma de explicar predições localmente, elas funcionam a partir de testes com alterações no conjunto de dados de entrada para identificar os impactos se outras

características apresentadas ao modelo fosse o caso. Em outras palavras, são feitas investigações através da simulação de cenários distintos. Molnar, Casalicchio e Bischl (2020) comentam que esse tipo de abordagem é encontrado em estudos de ciências sociais.

Por outro lado, a explicabilidade global visa apresentar o comportamento de um modelo de uma forma mais geral. A partir das predições observadas em diversos dados de entrada são extraídas métricas que podem indicar quais são as informações mais relevantes, como importância e efeito de variáveis (MOLNAR; CASALICCHIO; BISCHL, 2020). Wei, Lu e Song (2015) fizeram um estudo sobre a relevância de se analisar a importância das variáveis de um modelo e quais são as métricas e técnicas mais utilizadas. Com uma visão mais geral sobre o comportamento do modelo podemos identificar e discutir se as relações apontadas pelo modelo são pertinentes.

É possível obter interpretação global a partir da combinação de explicações locais de diversas predições (MESSALAS; KANELLOPOULOS; MAKRIS, 2019). No próximo capítulo será explorada uma das técnicas mais utilizadas de interpretação de modelos de aprendizado de máquina, a de obtenção de valores *SHAP*.

CAPÍTULO

4

VALORES SHAP

4.1 Metodologia

A expressão *SHAP* é uma abreviação de *SHapley Additive exPlanations* e foi apresentada por Lundberg e Lee (2017) para denominar uma metodologia que tem sido bastante utilizada para explicar o funcionamento de um modelo desenvolvido com aprendizado de máquina. Essa metodologia tem como objetivo criar um modelo mais simplificado e capaz de apresentar o comportamento geral do modelo complexo, para fazer isso são aplicados conceitos da área de teoria dos jogos. Essa área é considerado um campo da matemática que busca compreender como escolhas realizadas por pessoas influenciam outras escolhas e teve como expoente *Lloyd Stowell Shapley*, vencedor do *nobel* de economia de 2012 (ROTH, 2016). *Lloyd Shapley* foi quem criou o conceito de valor *Shapley*, que é uma forma de medir a contribuição de cada jogador para o resultado final de um jogo cooperativo e serviu como inspiração para o trabalho do Lundberg e Lee (2017).

No *SHAP* o processo de predição do modelo é apresentado como um jogo e cada variável do modelo é enxergada como um "jogador". Desta forma o valor Shapley é estimado através da contribuição de cada "jogador" diante da atuação dos demais, o que possibilita captar os comportamentos não lineares dos modelos complexos. Nessa abordagem é possível ter uma compreensão geral do comportamento de cada variável através da média de suas contribuições nas predições, além de viabilizar também avaliar a contribuição de cada variável para a predição de uma observação específica (RODRÍGUEZ-PÉREZ; BAJORATH, 2019).

4.2 Fundamentação teórica

De acordo com Lundberg e Lee (2017) considerando que F é o conjunto de todas as covariáveis envolvidas no modelo complexo e S é um subconjunto de F ($S \subseteq F$) a metodologia *SHAP* consiste em retreinar o modelo (simplificado) várias vezes e comparar o efeito da variável

nos modelos com e sem ela, possibilitando trazer uma métrica de importância da variável para esse modelo. Considerando que $f_{S\cup\{i\}}$ é o modelo com a variável envolvida e que f_S é o que foi retreinado sem a variável, as predições desses dois modelos são comparadas ($f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)$) para se ter uma ideia se a variável i é mais ou menos relevante para a performance do modelo complexo. Lundberg e Lee (2017) também afirmam que como a importância de cada variável depende da combinação com as demais esse cálculo deve ser feito para todos os subconjuntos S possíveis ($S \subseteq F\backslash\{i\}$). Dessa forma, o valor *SHAP* de cada variável pode ser calculado através da média ponderada de todas as diferenças possíveis, conforme a equação abaixo:

$$\phi_i = \sum_{S\subseteq F\backslash\{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!}[f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)] \tag{4.1}$$

Os autores ainda indicam que o cálculo dos valores *SHAP* pode ser feito através da equação 4.1 aplicada a amostras, isso reduz consideravelmente a quantidade de cálculos e modelos necessários para se estimar a participação de cada variável.

Lundberg e Lee (2017) também apontam algumas características dos valores *SHAP*, como: acurácia local, o que implica que o modelo simplificado construído para um conjunto de dados específico terá a mesma predição que o modelo original nesses dados; ausência, se o valor *SHAP* para uma dada variável é 0 significa que ela não participa na predição do modelo original; consistência, indica que variações na predição de um modelo devem estar compatíveis com as variações nas contribuições calculadas para cada variável; aditividade, a soma dos valores *SHAP* calculados para todas as variáveis com a média das previsões do modelo (valor de base do modelo) deve resultar no valor da predição do modelo original.

Em outras palavras, os valores *SHAP* são uma métrica que indica a contribuição de cada variável no resultado de um modelo considerando a relação com as demais. As estimações dos valores *SHAP* obtidas em uma predição de um modelo vão indicar o quanto a característica de cada variável (o valor dela) aumentou ou diminuiu essa predição, especificamente. Em modelos desenvolvidos com aprendizados de máquina e que identificam relações muito complexas essa métrica é muito útil, como será apresentado nesse trabalho.

4.3 Variações na implementação

A metodologia dos valores *SHAP* é agnóstica, ou seja, ela relaciona os dados de entradas com a saída de um modelo sem precisar analisar o modelo em questão. No entanto, o cálculo das contribuições das variáveis pode demandar muito esforço computacional. Diante disso, algumas das implementações dessa metodologia são direcionadas a alguns tipos de modelos específicos para poder otimizar a estimação dos valores *SHAP* (LUNDBERG; ERION; LEE, 2018).

De acordo com o que já foi exposto, o cálculo do valor *SHAP* de uma variável no modelo envolve comparar as predições em duas situações distintas, uma com e sem a participação dela. A maneira que essa comparação é realizada pode ser feita a partir de diferentes técnicas, por isso existem variações das implementações (LUNDBERG; LEE, 2017). É interessante destacar que, nessa metodologia, a estimação dos valores que permitem a interpretação de um modelo complexo de aprendizado de máquina é realizada através de outros modelos complexos de aprendizado de máquina.

Lundberg e Lee (2017) ao proporem a metodologia dos valores *Shap* apresentaram também algumas variações, como: *Kernel Shap*, que pode criar explicações de forma agnóstica e utiliza conceitos de explicações locais através de modelos lineares apontados pela metodologia *Lime*; o *linear Shap*, desenvolvido para modelos lineares e que estima os valores *Shap* através dos coeficientes desses modelos; *Deep Shap*, voltado para modelos de redes neurais profundas. No ano seguinte foi apresentada uma variação direcionada para modelos de *ensemble* de árvores, o *Tree Explainer*, que otimiza o processo de estimação dos valores *Shap* especificamente para esses modelos (LUNDBERG; ERION; LEE, 2018). Como neste trabalho serão utilizados modelos desenvolvidos com impulsionamento de árvores de decisão (*Boosting*), o *Tree Explainer* será a implementação utilizada.

4.4 Visualizações

Para facilitar uma melhor compreensão do comportamento de um modelo, algumas técnicas de visualização dos valores *SHAP* são empregadas. Esta seção apresenta alguns dos gráficos mais comuns para demonstrar os valores que foram estimados. Molnar (2021) mantém um *site* com alguns exemplos de gráficos que serão introduzidos nesse trabalho. O exemplo a seguir mostra a aplicação de um modelo com o algoritmo de floresta aleatória que contém 100 árvores em um conjunto de dados de 858 mulheres que apresentaram ou não câncer de colo de útero (variável resposta)[1], portanto um modelo de classificação binária, e suas respectivas características relacionadas à saúde (variáveis preditoras).

4.4.1 Gráfico de importância

O gráfico de importância, ou *Feature importance*, é um gráfico que busca trazer uma compreensão global do comportamento do modelo. Nele, os valores *SHAP* estimados para várias predições são consideradas para se obter uma métrica do quanto uma variável é importante no funcionamento desse modelo neste conjunto de predições. Essa métrica consiste em calcular a média do absoluto dos valores SHAP, ou seja, a média do impacto de cada variável na predição do modelo, desconsiderando se esse impacto foi para aumentar ou diminuir o valor predito (positivo ou negativo).

[1] Dados disponíveis em: https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29

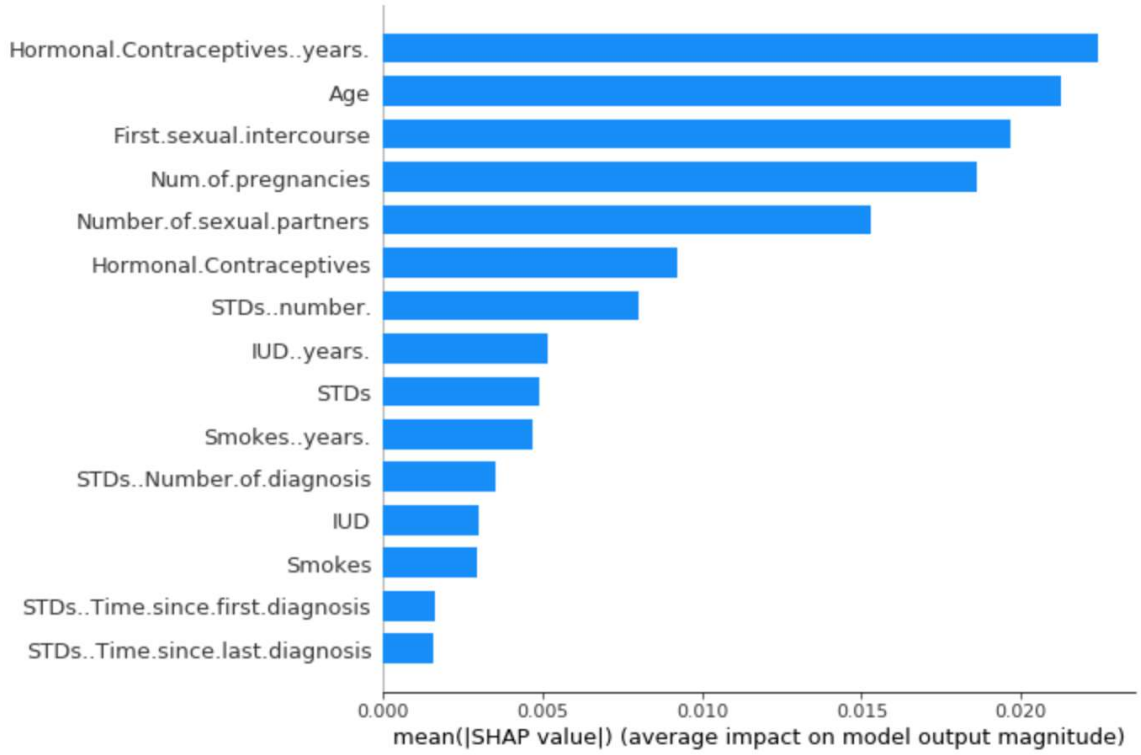

Figura 1 – Gráfico de importância. Fonte: (MOLNAR, 2021)

No gráfico 1 são apresentadas as variáveis que mais afetam a predição do modelo para o risco de câncer de colo de útero no conjunto de dados considerado. Nele, podemos mensurar se uma variável impactou, em média, muito mais do que outra. No entanto, não é possível analisar, através deste gráfico, se há alguma relação do valor de cada variável com o seu impacto. Para esse exemplo é apontado que o número de anos com uso de contraceptivos hormonais é o que mais altera o risco indicado pelo modelo de câncer de colo de útero, seguido da idade. É apresentado também que, em média, o número de anos com uso de contraceptivos hormonais altera quase 2,5% as probabilidades preditas pelo modelo e idade um pouco mais de 2%. Porém, não é indicado como ocorrem esses efeitos, isto é, se um maior tempo de uso de contraceptivo diminui ou aumenta o risco que é apontado pelo modelo, por exemplo. Para visualizar melhor esse comportamento é preciso produzir outro tipo de gráfico.

4.4.2 Gráfico de resumo

O gráfico de resumo, ou *Summary plot*, busca apresentar os efeitos de cada variável no funcionamento do modelo. Para a amostra indicada são levantadas as variáveis que mais impactaram e como foi o impacto.

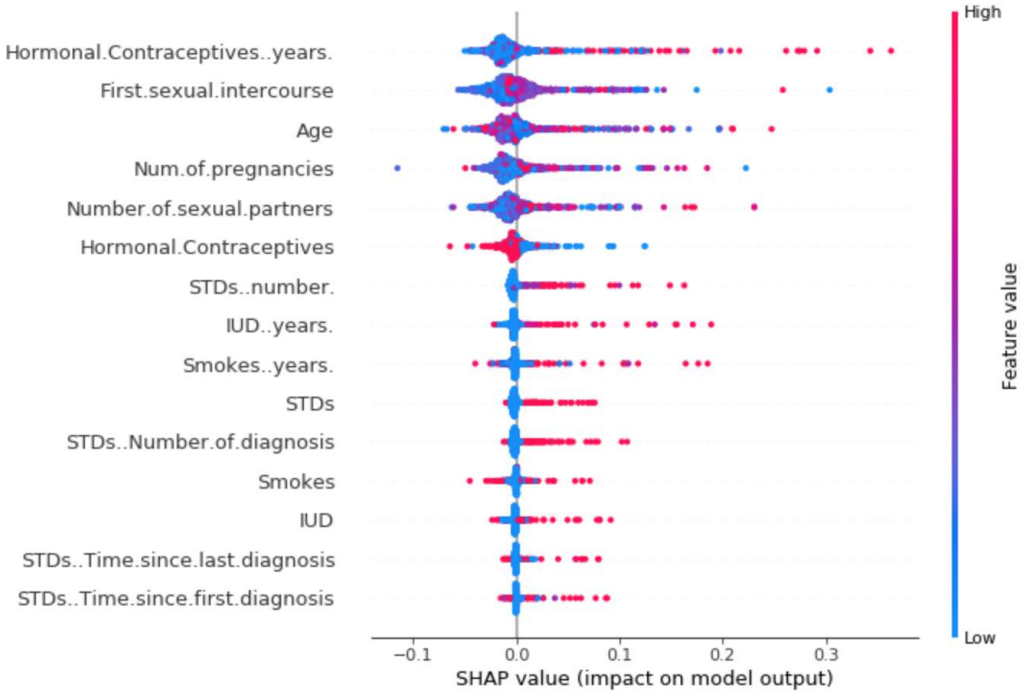

Figura 2 – Gráfico de resumo. Fonte: (MOLNAR, 2021)

Neste gráfico, cada ponto é o valor *SHAP* estimado de uma variável para uma das mulheres presentes nesse conjunto de dados. Podemos entender um pouco da distribuição dessas contribuições através da concentração desses pontos. A escala de cor indica se o valor da variável, nessa amostra, é alto ou baixo. Com isso, é possível identificar alguns dos padrões desse modelo complexo, como por exemplo: maior tempo de uso de contraceptivos, em anos, contribuem para uma predição do modelo de maior risco de câncer de colo de útero. Molnar (2021) enfatiza que não necessariamente o comportamento do modelo implica em relação causal, ou seja, é preciso avaliar mais estudos para revelar se o uso de métodos contraceptivos por muitos anos causa um maior risco, o que o modelo indica são padrões presentes nos dados. Além disso, é possível observar também que em outras variáveis não há uma relação clara entre valores maiores ou menores com uma predição de risco maior ou menor (pontos com cores distintas, ou seja, que têm valores da variável diferentes apresentando efeitos semelhantes no risco indicado pelo modelo), o que evidencia um comportamento não linear.

4.4.3 Gráfico de dependência

Para investigar melhor o efeito de uma variável na predição de um modelo há o gráfico de dependência, que apresenta os valores *SHAP* estimados no conjunto de dados em questão de acordo com o valor da variável original. Através desse tipo de gráfico pode ser observado se há algum padrão a partir de um valor específico da variável.

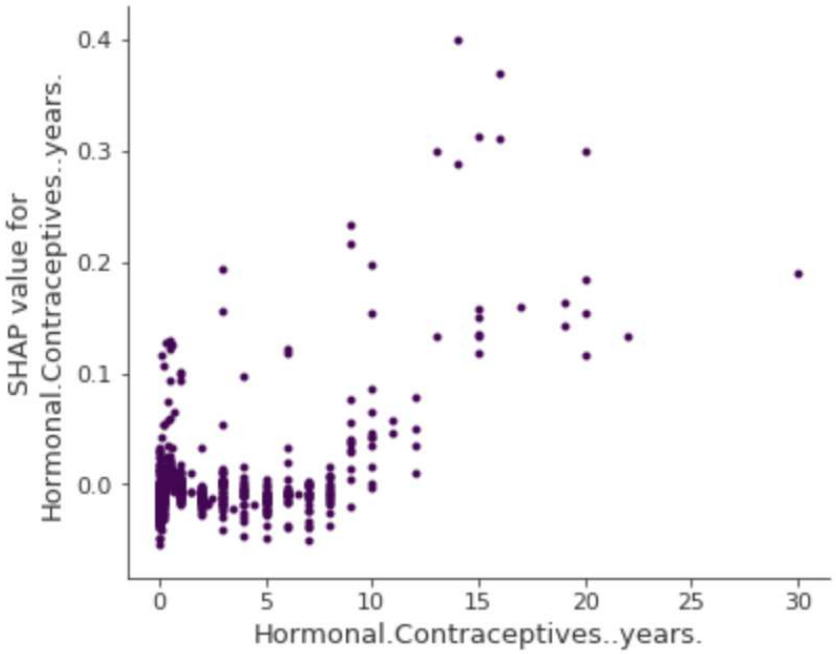

Figura 3 – Gráfico de dependência. Fonte: (MOLNAR, 2021)

Assim como no gráfico de resumo, cada ponto aqui representa a quantidade de anos de uso de contraceptivo e o valor *SHAP* estimado para cada uma das mulheres presentes nesse conjunto de dados, mas aqui temos a informação mais detalhada do valor da variável em questão. Observamos na figura 3 que a partir de 9 anos de uso de contraceptivo há uma tendência de crescimento do risco de câncer, indicando que quanto maior o uso desse método maior a propensão da mulher desenvolver essa doença.

4.4.4 Gráfico de força

Uma das principais características da metodologia de interpretabilidade de modelos *SHAP* é a aditividade, ou seja, a soma dos valores *SHAP* estimados para uma predição específica resulta na diferença entre a predição do modelo e a média geral das predições. Com isso, uma das ferramentas de visualização é o gráfico de força, ou *force plot*. Esse gráfico consiste em apresentar os desvios que cada variável provocou na composição do valor final de saída do modelo, ou seja, os valores *SHAP* para cada variável e o resultante desses valores. Como esses valores são referentes a uma predição específica esse gráfico é muito útil para compreender o funcionamento local do modelo. Uma variável pode contribuir em um menor ou maior grau, de acordo com as demais características e forma que o modelo funciona para predições próximas.

Molnar (2021) trouxe como exemplo 2 gráficos de força para o modelo que indica o risco de câncer de colo de útero. São duas mulheres, portanto, com características e riscos de câncer distintos, como cada característica contribuiu no aumento ou diminuição do risco pode ser visualizado através desses gráficos.

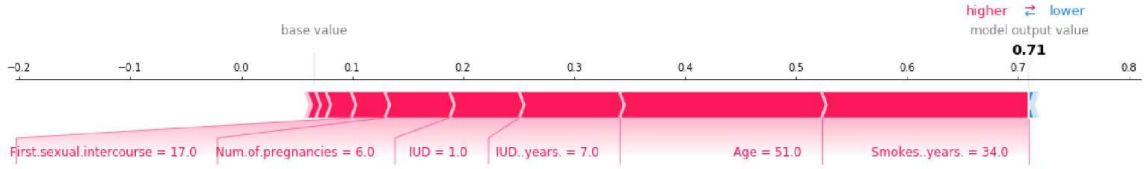

Figura 4 – Gráfico de dependência. Fonte: (MOLNAR, 2021)

No primeiro caso indicado na figura 4 o risco de câncer é elevado, uma probabilidade indicada pelo modelo de 71%, e praticamente nenhuma característica dessa mulher diminuiu o risco da doença. O fato dela ter fumado por 34 anos e ter 51 anos são as características que mais impactaram para o aumento desse risco. Vale notar o *base value* presente no gráfico de força, ele indica o valor médio das predições desse modelo. Como mencionado anteriormente, a soma dos valores *SHAP* vão indicar a diferença entre o valor específico de risco dessa predição e o *base value*.

Figura 5 – Gráfico de dependência. Fonte: (MOLNAR, 2021)

Já no exemplo da imagem 5, o risco apontado pelo modelo dessa mulher ter câncer de colo de útero é menor, 6%. Existem características que diminuíram o risco, como 7 anos de uso de contraceptivos e 42 anos de idade. No entanto, ela já foi diagnosticada 2 vezes com doença sexualmente transmissível (*STD - Sexually Transmitted Diseases*) e ficou 6 anos com dispositivo intra uterino (*IUD - intrauterine device*), características que aumentaram um pouco o risco de câncer. Nos capítulos seguintes a metodologia de valores *SHAP* e as formas de visualizar essas estimações serão aplicadas e discutidas em conjuntos de dados de outros contextos.

CAPÍTULO

5

ENEM

5.1 Sobre o ENEM

O Exame Nacional do Ensino Médio, ENEM, é uma prova que foi criada pelo governo federal brasileiro em 1998 com o objetivo de avaliar os conhecimentos de alunos brasileiros que terminaram o segundo grau (SILVEIRA; BARBOSA; SILVA, 2015). De acordo com Sakalauskas e Trevisan (2017) em 2009 esse exame foi reformulado e passou a ser uma das formas de acesso dos alunos às universidades, através do SiSU - Sistema de Seleção Unificada. Além disso, os autores apontam que a partir desse mesmo ano o exame passou a ser utilizado também como meio de obtenção do certificado de conclusão do ensino médio.

Para cada edição da prova o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (2021) - INEP, autarquia responsável pela organização do exame, disponibiliza, de forma anonimizada, os microdados do exame. Nele há informações socieconômicas dos candidatos, assim como o desempenho.

Neste trabalho foram considerados os microdados da edição de 2021, que é a mais recente disponível no site do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (2021) em março de 2023. No conjunto de dados há o registro de 3,389,832 candidatos, com dados informados pelo participante sobre escolaridade e atividade profissional dos pais, renda mensal da família, acesso a alguns bens e serviços específicos, a nota do exame em cada prova, cor da pele, sexo, entre outros.

5.2 Ajustando um modelo

Nesta seção serão descritas as etapas que foram realizadas para ajustar um modelo de aprendizado de máquina.

5.2.1 Objetivo

Segundo Almeida (2020) a partir de 2017 o exame nacional de ensino médio deixou de ser utilizado como meio de obtenção do certificado de conclusão do ensino médio. No entanto, essa prova tem a característica de utilizar a *TRI*, teoria de resposta ao item, para padronizar as notas dos candidatos entre os anos. Neste trabalho serão aplicados os critérios que eram utilizados como requisitos para emissão do certificado de conclusão do ensino médio na definição se os participantes alcançaram um desempenho mínimo no exame.

O critério de emissão desse certificado era de nota mínima de 450 pontos em cada uma das 4 provas objetivas e 500 na redação. Considerando esse critério, foram identificados os participantes que atingiram essas notas e ajustado um modelo de classificação. O objetivo é identificar se existem padrões nas informações dos participantes que indicam se eles são mais ou menos propensos a ter um desempenho mínimo na prova.

5.3 Preparo dos dados

Para poder desenvolver um modelo de aprendizado de máquina é necessário estruturar o conjunto de dados que vai servir como base para o seu ajuste. Nos dados do Enem há muitas variáveis que são categóricas, para a maioria delas foi utilizada uma técnica chamada *One hot encoding*, que busca transformar essas variáveis categóricas em variáveis numéricas. Para cada categoria é criada uma nova variável que contém os valores 0 e 1, onde 1 indica se a observação específica pertence à categoria que essa variável representa. A variável sexo nesse conjunto de dados, por exemplo, que está registrada como texto e tem as marcações “F” para feminino e “M” para masculino deu origem a variável sexo_M onde o número 1 indica que o participante é do sexo masculino e 0 do feminino.

Há também variáveis que são categóricas mas apresentam ordem entre as categorias. A resposta do participante para a pergunta “Qual é a renda mensal de sua família? (Some a sua renda com a dos seus familiares.)”, por exemplo, é uma das alternativas entre as letras A e Q, onde A indica que não há nenhuma renda, H que a renda é entre R$4.400,00 e R$5.500,00 e Q acima de R$ 22.000,00 (cada letra indica uma faixa). Para essas variáveis foram criadas variáveis correspondentes com números, onde 0 representa a menor faixa e quanto maior o número, maior a faixa. Em algumas das perguntas do questionário socio-econômico havia a alternativa que indicava desconhecimento, como a de escolaridade do pai, em que a letra H indica que o participante não sabe a resposta. Para esse tipo de pergunta foi identificado qual que é o ordenamento das alternativas e avaliado onde o desconhecimento da resposta se encaixa, no caso da escolaridade do pai a letra A aponta que ele nunca estudou e a letra G que ele completou a pós-graduação, na variável numérica que indica a escolaridade do pai os participantes que não sabiam a escolaridade do pai ficaram com o valor 0, os que o pai nunca estudou 1 e assim por diante até o valor 7, que indica que o pai terminou a pós graduação.

Esses tratamentos de dados são necessários porque a maior parte dos algoritmos de aprendizado de máquina demanda que os dados sejam representados numericamente. Além disso, quando existe relação de ordem nas informações que uma variável representa é importante manter essa ordem na representação numérica. Algumas variáveis foram retiradas do desenvolvimento do modelo por não serem muito relevantes no estudo, como por exemplo as respostas das perguntas “Na sua casa tem aparelho de DVD?” e “Na sua residência tem aspirador de pó?”. Ao final, realizando todo o pré-tratamento, o conjunto de dados disponível para desenvolvimento do modelo continha 38 variáveis.

A variável objetivo foi construída a partir das notas dos participantes. Como mencionado o critério de marcação de desempenho mínimo foi de pelo menos 450 pontos em cada uma das provas objetivas e 500 pontos na redação. Logicamente, as notas nessas provas não participaram como variável preditora. Dos 3,389,832 inscritos, pouco mais de 2 milhões e 200 mil (mais especificamente 2,238,107) compareceram nos dois dias de prova. Entre os que compareceram, 1,270,935 não conseguiram atingir todas as notas mínimas apontadas, representando quase 57%, o que revela que a maior parte tem um desempenho insatisfatório.

5.4 Ajuste do modelo

O algoritmo utilizado para ajustar esse modelo foi o *lightgbm classifier*, que implementa técnicas de impulsionamento de árvores de decisão (*boosting*) e é bastante utilizado. No conjunto de dados foi feita uma amostragem aleatória em que 70% dos registros foram separados para desenvolver o modelo e 30% para avaliar o desempenho. Para ajustar os hiperparâmetros do modelo foi utilizado o algoritmo *OPTUNA*, buscando otimizar através de validação cruzada *k-fold* entre os registros disponíveis para desenvolvimento, nesse caso foram utilizados 5 *folds*. O objetivo definido foi de maximizar a métrica *AUC - Area under the ROC curve*, que é amplamente utilizada em contextos de classificação e quanto maior a métrica melhor a capacidade de discriminação do modelo.

Para esse estudo os hiperparâmetros utilizados foram: *n_estimators*, que indica a quantidade de árvores que o modelo apresentou; *learning_rate*, ou taxa de aprendizagem, que controla o quanto os pesos dos modelos serão atualizados em cada iteração; *max_depth*, que aponta o máximo de níveis que cada árvore pode atingir e *num_leaves*, que restringe o máximo de folhas (largura) das árvores.

Para o ajuste desse modelo os valores dos hiperparâmetros obtidos são apresentados na tabela 1:

Hiperparâmetro	Valor
n_estimator	200
learning_rate	0.0675
max_depth	44
num_leaves	150

Tabela 1 – Parâmetros do modelo de LightGBM para o ENEM.

A quantidade de arvores, assim como a profundidade e a largura máxima que elas podem ter, dão uma dimensão da complexidade desse modelo e evidenciam que entender o seu comportamento a partir de sua arquitetura é algo extremamente difícil. Para esse modelo, especificamente, são envolvidas 200 árvores (*n_estimator*), onde cada uma delas pode apresentar até 44 níveis (*max_depth*) e 150 folhas (*num_leaves*), apontando a inviabilidade de um ser humano acompanhar o cálculo de uma predição. Na próxima seção será brevemente apresentada a taxa de acerto do modelo e na seguinte como foi utilizada a metodologia de valores *SHAP* para buscar compreender o seu funcionamento.

5.4.1 Resultados

O conjunto teste é composto por 656,989 registros de candidatos da prova e que não participaram no desenvolvimento do modelo. Na tabela 2 é apresentada a matriz de confusão. Foi chamado de aprovado o participante da prova que conseguiu ter a nota maior do que o mínimo que foi mencionado.

Tabela 2 – Matriz de confusão do modelo de classificação do Enem

		Classe prevista	
		Reprovado	Aprovado
Classe real	Reprovado	294,771 (45%)	79,024 (12%)
	Aprovado	114,355 (17%)	168,839 (26%)

A taxa de acerto desse modelo foi de aproximadamente 71 %.

5.5 Interpretação

No gráfico de importância 6 é possível visualizar quais variáveis foram mais relevantes para esse modelo.

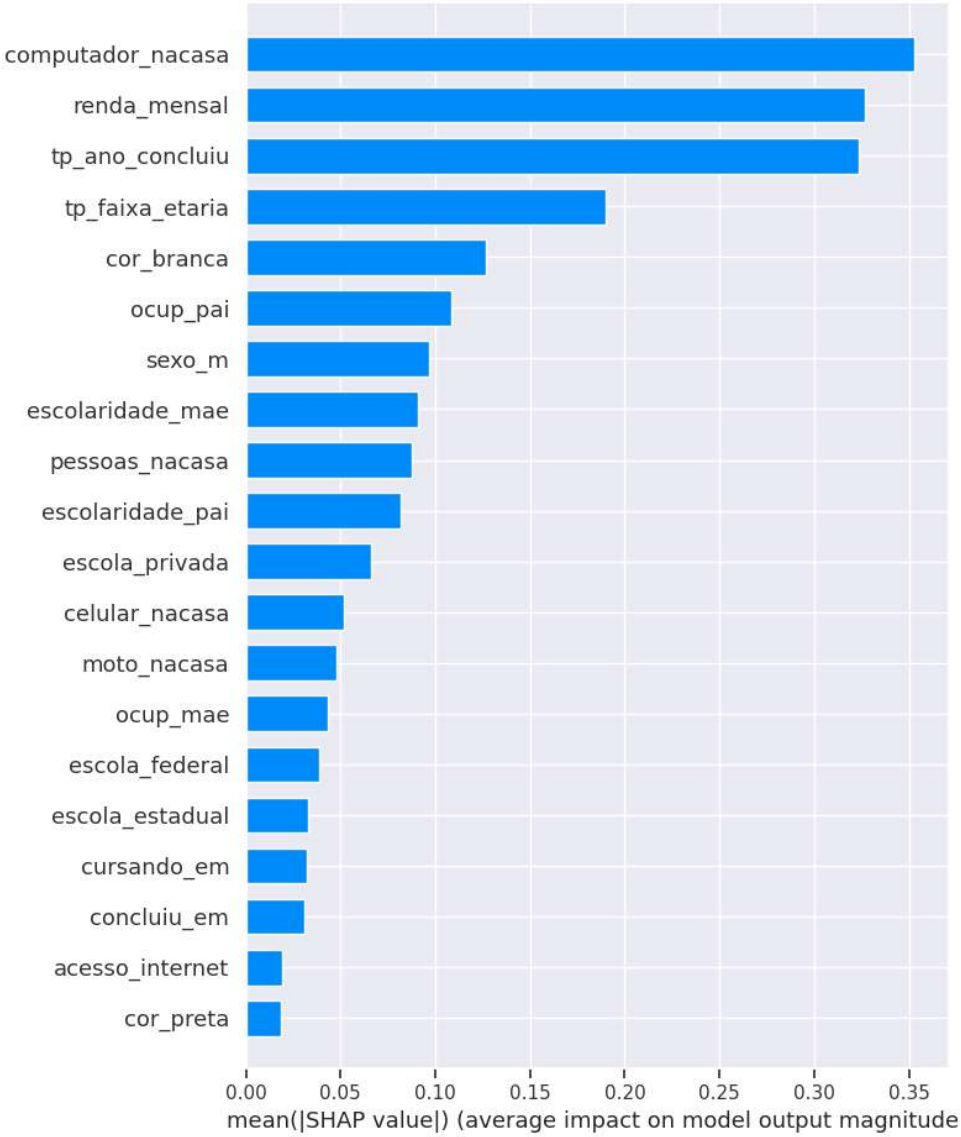

Figura 6 – Gráfico de importância

É interessante notar que a variável *computador_nacasa* é a que apresentou maior importância nesse modelo. 2021 foi um ano atípico, muito impactado pela pandemia, quando o ensino era predominante remoto. Renda mensal da família também é uma característica que obteve destaque na predição se o candidato alcançaria a nota mínima no exame. Na figura 7 identificamos o sentido que essas variáveis participaram no modelo.

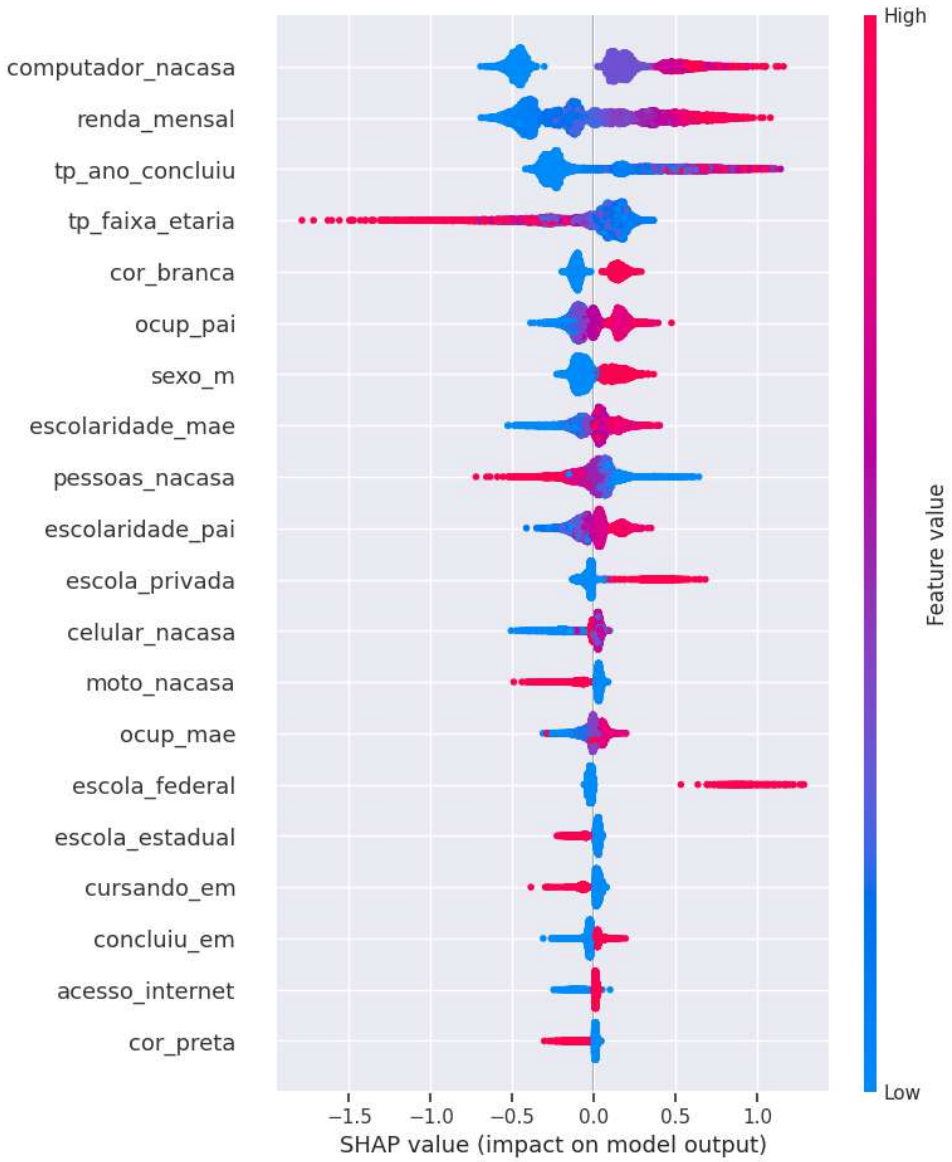

Figura 7 – Gráfico de resumo

Cada ponto indicado no gráfico da figura 7 representa a contribuição de uma variável na predição do modelo para um candidato, pontos com a cor mais próxima ao vermelho indicam um maior valor da variável em questão, enquanto que mais próxima ao azul um valor menor. Dessa forma, conseguimos visualizar que o fato do participante da prova ter um ou mais computadores na casa levam a uma predição do modelo de melhor desempenho. O mesmo vale para o contrário, se o candidato não tiver um computador na casa ele é mais propenso a ter uma nota menor que o mínimo considerado.

Na figura 7 conseguimos identificar alguns padrões. A escala do gráfico está em *logito* (não é a probabilidade), mas quanto mais à direita maior a contribuição da variável na probabilidade do candidato alcançar a nota mínima. De acordo com o modelo, o perfil dos candidatos que têm maior probabilidade de atingir a nota mínima inclui: i) alta renda familiar, ii) pais com ocupação mais valorizada, iii) brancos, e iv) homens. Por outro lado, o perfil dos que têm baixa

probabilidade de atingir a nota apresenta: i) ausência de computador na casa, ii) renda familiar baixa, e iii) mulheres.

É sempre importante lembrar que os padrões identificados pelo modelo não implicam, necessariamente, em causalidade. No entanto, é interessante identificar as variáveis mais relevantes e avaliar se esses padrões corroboram com outros estudos. A ocupação do pai se apresentou mais relevante que a da mãe, porém, a escolaridade da mãe é mais relevante que a do pai. Pode-se levantar a hipótese de que a mãe é mais presente diretamente na educação dos filhos, enquanto o pai seria mais responsável pela parte financeira. Nota-se na figura 7 que a cor branca se apresenta como fator importante, assim como a cor preta, mas em sentidos distintos. Há diversos estudos, como o da Bento (2005), que discutem a desigualdade entre brancos e negros e apontam a importância de políticas afirmativas como as de cotas. É revelado também que o fato do candidato estudar em uma escola privada contribui para uma maior probabilidade dele atingir a nota mínima. Além disso, entre os candidatos que estudam em escolas públicas, fica clara a distinção entre os alunos de escolas federais e estaduais, é apontado pelo modelo que as federais contribuem significativamente de forma positiva no desempenho dos seus alunos, enquanto que as estaduais impactam negativamente. Essa diferença entre os tipos de escolas na educação brasileira também é tópico relevante na literatura, o resultado apontado no artigo de Sampaio e Guimarães (2009), por exemplo, condiz com o que foi observado no gráfico.

As figuras 8, 9 e 10 apresentam os gráficos de força de 3 candidatos distintos. Nelas são destacadas quais características foram mais relevantes para o modelo identificar a probabilidade destes candidatos alcançarem uma boa nota no exame.

Figura 8 – Exemplo 1 de Gráfico de força

Para o candidato apresentado na Figura 8 a probabilidade de obtenção da nota mínima é bem baixa, 6% apenas. As características que mais impactaram para essa baixa probabilidade foram renda e computador na casa. Como ele tem uma renda familiar bem baixa e nenhum computador na casa em um período recente à pandemia, a projeção é de que ele não alcance uma boa nota no exame. Nota-se também outras informações que o colocaram como um perfil que tem notas baixas na prova, como: i) não ter concluído o ensino médio, ii) sua faixa etária estar na categoria 4 (19 anos) iii) a ocupação do pai se encaixar em uma área que é menos privilegiada, iv) ele não ser branco, e v) a mãe ter baixa escolaridade.

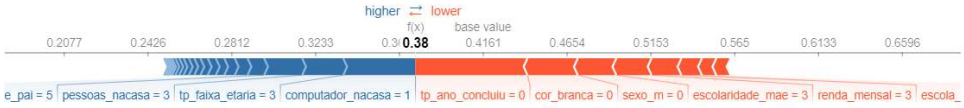

Figura 9 – Exemplo 2 de Gráfico de força

Já na figura 9 é apresentado uma candidata com probabilidade um pouco melhor de alcançar a nota mínima no exame, 38%. O fato de ter computador na casa, a faixa etária igual a 3 (18 anos) e poucas pessoas morando com ela contribuíram positivamente na estimativa da probabilidade, enquanto que não ter concluído o ensino médio, ser mulher e não branca impactaram de forma negativa.

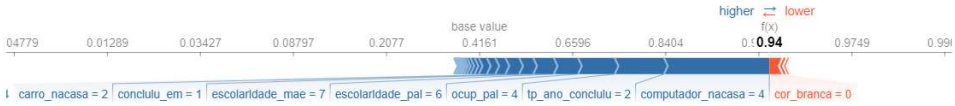

Figura 10 – Exemplo 3 de Gráfico de força

A figura 10 é referente a um candidato com alta propensão de alcançar a nota mínima no exame, 94%. Trata-se de um candidato que possui muitos computadores na casa, os pais têm um grau de escolaridade mais alto e o pai tem uma ocupação que é considerada como valorizada. A única característica que aparece impactando negativamente é a cor não branca.

5.6 Conclusão

O modelo pode ser utilizado para compreender melhor as características dos candidatos que têm um desempenho mínimo no Exame Nacional do Ensino Médio. Ele também pode revelar padrões que ocorrem em parcelas específicas da população brasileira. A interpretação de um modelo aplicado a dados socieconômicos pode servir de auxílio para fomentar estudos posteriores e investigar relações entre os dados que podem ou não serem causais. Além disso, pode-se mensurar quais são as características mais relevantes e relacionar essas informações com outros estudos para servir como auxílio na elaboração de políticas públicas mais efetivas.

CAPÍTULO

6

SCORE DE CRÉDITO

6.1 Concessão de crédito

O mercado de crédito é capaz de acelerar o crescimento econômico de um país, aumentar a geração de emprego e viabilizar aquisição de bens de consumo (OLIVEIRA, 2018). Porém, é necessário avaliar muito bem como o crédito é oferecido, de forma que a inadimplência se mantenha em níveis razoáveis e o ciclo de crédito seja viável. De acordo com Forti (2018), modelos de previsão cada vez mais acurados são empregados para reduzir as taxas de inadimplência, possibilitando que a concessão não ocorra para os proponentes a tomador de crédito mais propensos a inadimplir. Para esse estudo de caso, foram levantadas informações sobre as características de empresas no momento que elas obtiveram crédito e se houve ou não inadimplência posteriormente.

6.2 Dados utilizados

Foram obtidos dados reais de uma cooperativa que oferece crédito para pessoas jurídicas. Esses dados trazem informações cadastrais, como idade e porte, e de comportamentos financeiros da empresa, como valor de restrições (dívidas), tempo que ela apresentou essas restrições, taxa de pagamento das dívidas e consultas para acesso a crédito. Esse conjunto de dados contém o registro de 58,000 empresas que obtiveram crédito junto à cooperativa e mais de 2 mil possíveis variáveis explicativas, além da informação se houve ou não inadimplência (variável resposta). O percentual de registros apontados como inadimplentes na variável resposta, nesse caso, foi de 7%. Devido à sensibilidade desses dados o nome da cooperativa e detalhes das variáveis envolvidas não serão apresentados.

6.3 Objetivo do modelo

O modelo tem como propósito relacionar os dados para identificar os padrões das empresas que cumprem e não cumprem os eventuais empréstimos que podem ser obtidos. Uma vez identificados esses padrões é possível criar um *score* que, baseado na probabilidade estimada, é capaz de predizer a propensão de uma empresa pagar o empréstimo. No estudo, uma probabilidade próxima de 1 indica que a empresa é altamente propensa e 0 que a empresa muito provavelmente vai inadimplir. A partir desse *score* a cooperativa pode elaborar uma política de crédito, que são diretrizes para decidir se a empresa, dadas as suas características e score, vai ter acesso ao crédito ou não.

6.4 Ajuste do modelo

O conjunto de dados foi dividido aleatoriamente em dois grupos, o conjunto de desenvolvimento, que representa 70% dos registros e onde foi feito o ajuste do modelo, e conjunto validação, onde as métricas de avaliação foram obtidas. O algoritmo utilizado foi o *lightgbm* e a busca pelos hiperparâmetros foi realizada através do *optuna*. Para esse conjunto, que possui uma quantidade de registros menor, foi adicionado um hiperparâmetro chamado *num_child_samples* que controla a quantidade mínima de registros em um nó das árvores do modelo, podendo auxiliar no controle do super ajuste. Os valores dos hiperparâmetros encontrados são indicados na tabela 3.

Hiperparâmetro	Valor
n_estimator	93
learning_rate	0.0427
max_depth	22
num_leaves	38
num_child_samples	145

Tabela 3 – Parâmetros do modelo de concessão de crédito.

Este modelo apresentou 93 árvores (*n_estimators*), que podiam ter profundidade máxima de 22 níveis (*max_depth*) e com 38 folhas (*num_leaves*) que continham pelo menos 145 registros (*num_child_samples*) durante o seu ajuste.

6.5 Resultados

Uma métrica de avaliação de modelo amplamente utilizada no mercado de crédito é a estatística *Kolmogorov-Smirnov* (KS), que é uma medida entre as distribuições acumuladas dos adimplentes e dos inadimplentes, e quanto maior melhor o modelo discrimina entre as duas classes (PICCIN, 2022). Além disso, foram levantados os percentuais de inadimplência por faixa das probabilidades apontadas pelo modelo.

Na imagem 11 é possível visualizar a distribuição das empresas inadimplentes e adimplentes do conjunto de validação. As empresas inadimplentes se concentram nas probabilidades mais baixas de pagamento, enquanto as adimplentes nas mais altas. Além disso, o ponto em que é obtido o máximo de diferença entre essas duas distribuições é destacado através da linha vertical pontilhada, é nesse ponto que a métrica de KS é obtida. Para esse modelo, o KS obtido nesse modelo foi de 41,5%.

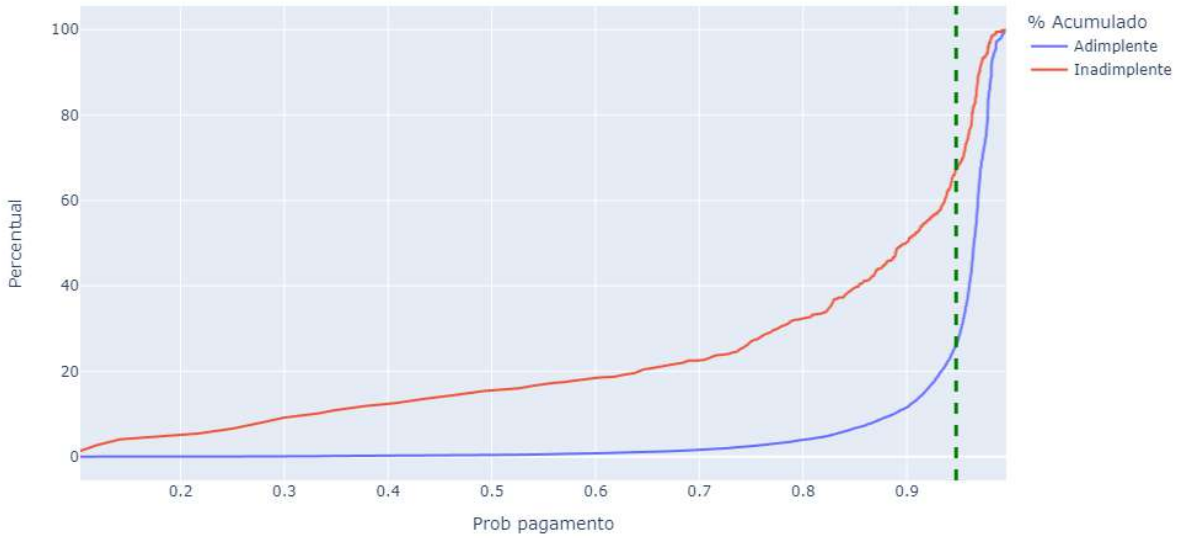

Figura 11 – Distribuição acumulada pela Probabilidade do Modelo

Na figura 12 foram definidas 10 faixas de probabilidade no conjunto de validação, de forma que a quantidade de empresas em cada uma delas fosse muito próxima (cada grupo contém aproximadamente 10% dos registros disponíveis na validação). É apontado que o grupo mais arriscado apresenta uma taxa de 29% de inadimplência, enquanto o menos arriscado apenas 1%. Em outras palavras, considerando que, no geral, 7% das empresas não pagam o empréstimo obtido, através desse modelo é possível separar em grupos em que as taxas são muito diferentes.

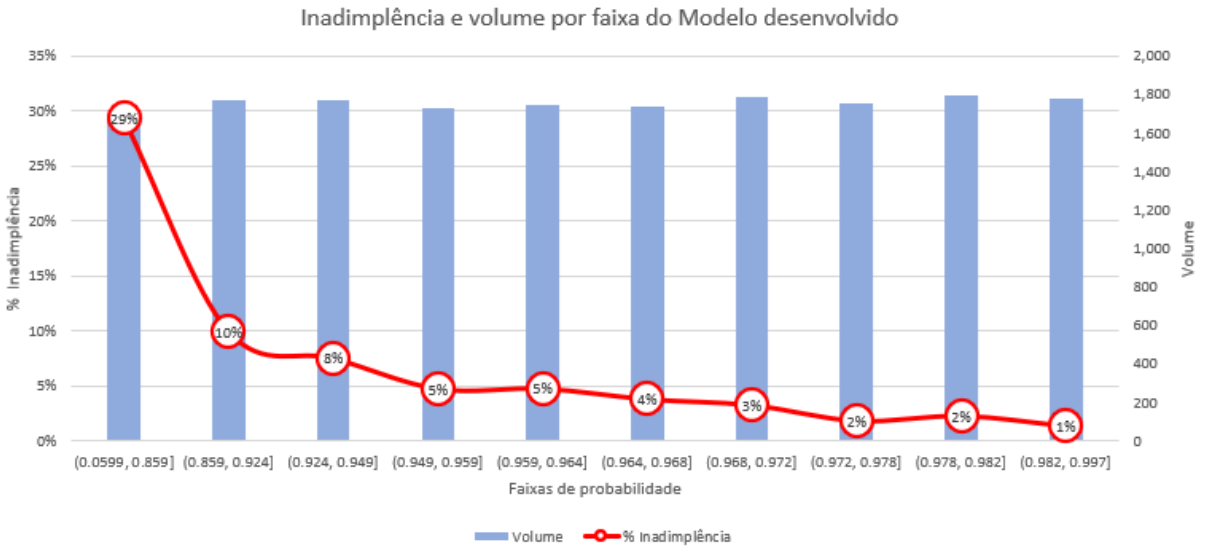

Figura 12 – Taxa de inadimplência por faixa

6.6 Interpretação

No gráfico de importância da figura 13 são indicadas as 20 variáveis mais relevantes nesse modelo. Os nomes das variáveis foram alterados para não indicar características sensíveis, mas, de forma simplificada, os dados trazem informações sobre valor de dívida, tempo de algum evento relacionado a consulta ou dívida, idade da empresa e taxa de pagamento das dívidas. É possível observar que a informação mais relevante é sobre valor e tempo de dívida.

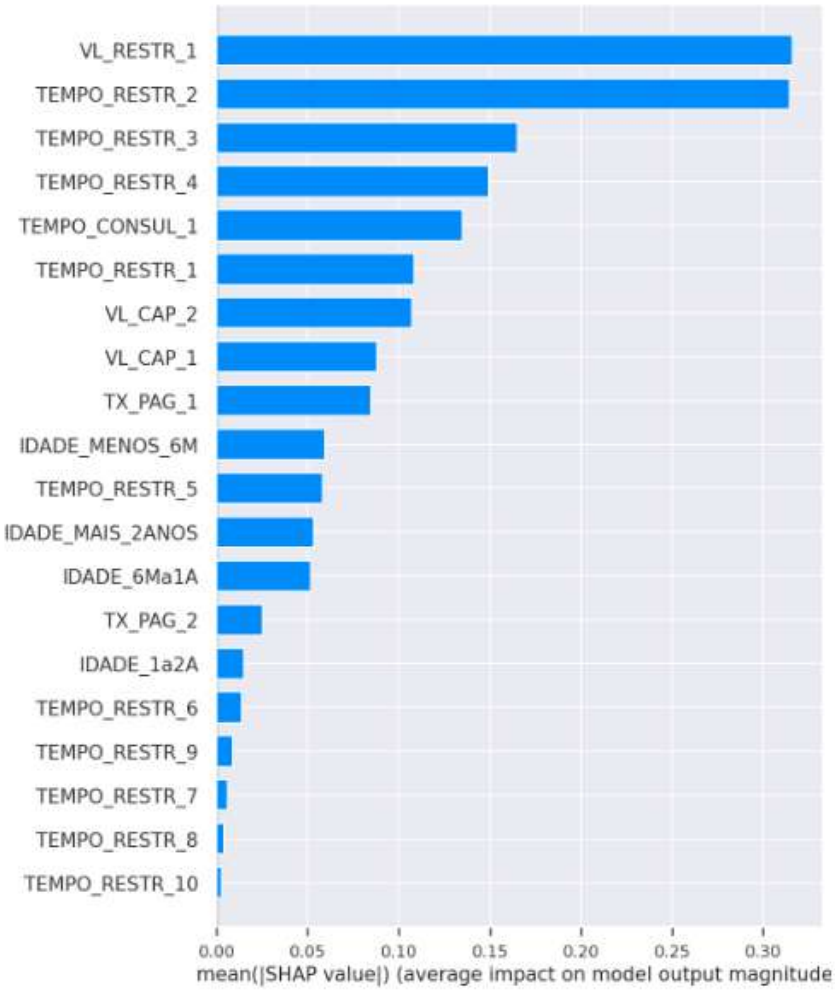

Figura 13 – Gráfico de importância

No gráfico de resumo da figura 14 são demonstradas as 10 variáveis mais importantes, sendo possível verificar os sentidos dos impactos dessas informações na probabilidade estimada de uma empresa inadimplir. O que chama a atenção nesse gráfico é a variável *IDADE_MENOS_-6M*, uma variável indicadora que aponta se a empresa é recém fundada (menos de 6 meses de fundação) ou não. Pelo que aponta o gráfico, empresas muito novas têm um risco de inadimplência associado menor que as demais (quanto mais à direita, mais a contribuição da característica para uma probabilidade estimada menor de inadimplência). Esse tipo de interpretação não está alinhada com as expectativas, o que normalmente é visto é que empresas recém constituídas que estão em busca de crédito não estão consolidadas no mercado e apresentam um risco maior de inadimplência.

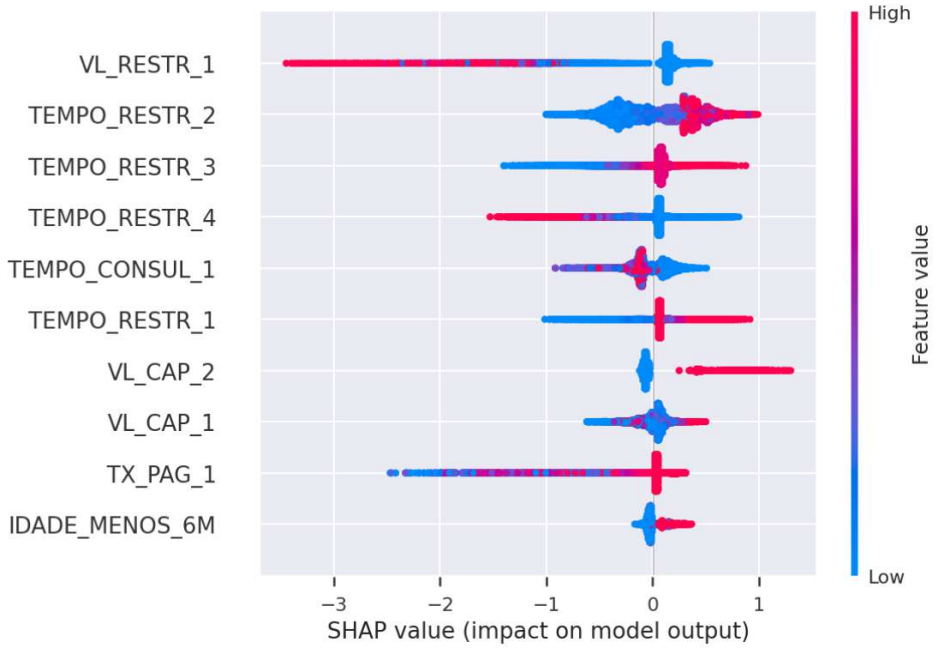

Figura 14 – Gráfico de resumo

Os gráficos de *boxplot* da figura 15 foram feitos com os dados de validação e indicam as distribuições das probabilidades estimadas do crédito ser pago (Prob Pagamento) das empresas, de acordo com a faixa de idade.

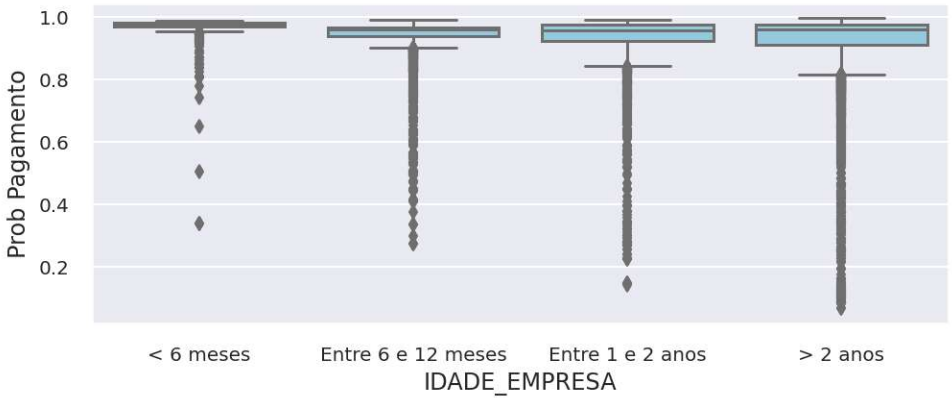

Figura 15 – Boxplot das probabilidades por idade

É observado que o modelo realmente indica que essas empresas recém constituídas são menos arriscadas (O *boxplot* aponta maior concentração nas probabilidades altas de pagamento). Vale notar que essas probabilidades estão relacionadas a baixa taxa de inadimplência presente no conjunto de dados (7%).

Para efeito de comparação, foi obtido o que se chama de *score* genérico, um indicador de propensão de pagamento que empresas de soluções de crédito oferecem. O score genérico possui esse nome porque é construído considerando amostras de diversos segmentos do mercado de concessão de crédito. No caso de crédito para pessoa física, por exemplo, a amostra utilizada para

o ajuste de um score genérico envolve crédito no varejo, financiamento através de financeiras e etc. Um procedimento comum é comparar o *score* genérico com a probabilidade de um modelo que foi ajustado para avaliar se esses modelos apresentam ou não comportamento similar.

A figura 16 apresenta os *boxplots* do *score* genérico nas mesmas empresas da figura 15, todas presentes no conjunto de validação. É possível observar que as empresas mais jovens têm uma distribuição de risco associada mais concentrada em valores mais baixos (o *score* é um número de 0 a 1000, onde números mais baixos indicam maior risco). Em outras palavras, há evidências de que o que foi apontado pelo modelo desenvolvido com os dados da cooperativa é diferente do que é normalmente observado no mercado.

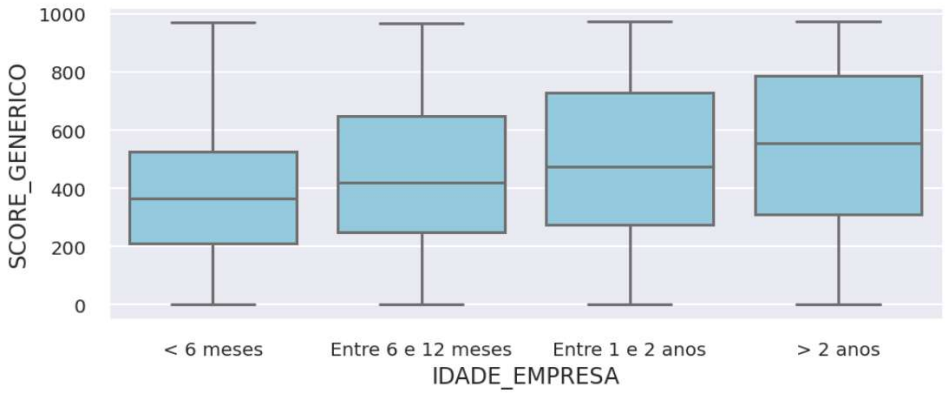

Figura 16 – Boxplot de um score genérico por idade

O gráfico de força 17 exibe um exemplo de empresa com menos de 6 meses que o modelo desenvolvido com os dados da cooperativa apresentou alta probabilidade de pagamento.

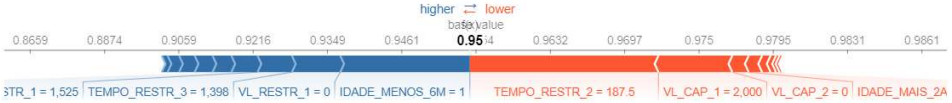

Figura 17 – Gráfico de força - Modelo desenvolvido com dados da cooperativa

Observa-se que o fato dessa empresa ser muito recente contribuiu para um menor risco associado (mais do que 1% de aumento na probabilidade de pagamento).

6.6.1 Buscando um modelo alternativo

Em muitos algoritmos de aprendizado de máquina há a implementação do que é chamado de restrição de monotonicidade. Através desse tipo de restrição é imposto que a atuação de uma determinada variável no modelo desenvolvido seja apenas em um sentido, isto é, conforme o valor da variável preditora sob efeito de restrição aumenta a contribuição dessa variável vai apenas aumentar (ou diminuir) a saída do modelo (SHARMA; WEHRHEIM, 2020). Este tipo de abordagem é bastante relevante quando se sabe o comportamento esperado entre duas variáveis

e/ou que a informação que está contida nos dados não é tão representativa em relação ao cenário em que o modelo vai ser utilizado. Ou seja, é uma maneira de interferir no aprendizado que o algoritmo vai realizar a partir dos dados.

Uma forma de buscar adequar o modelo de concessão de crédito que foi desenvolvido para a cooperativa foi de aplicar a restrição de monotonicidade na variável indicadora referente a idade da empresa. Neste caso a restrição foi negativa, ou seja, se uma empresa é muito recente (menos do que 6 meses) o valor da variável indicadora vai ser maior (vai ser 1) do que as demais empresas (0) e o algoritmo só vai incluir essa variável no modelo se ela atuar para diminuir a probabilidade de pagamento. O algoritmo *lightgbm* possui um parâmetro referente a esse tipo de restrição e nele é definido para cada variável preditora se a variável deve ou não ser monotônica e, se for, em qual o sentido.

Após desenvolvido esse novo modelo, com restrição apenas na variável IDADE_MENOS_6M, foi obtida a métrica de KS e a ordenação de risco de acordo com as faixas de probabilidade desse modelo. Na imagem 18 são indicadas as distribuições acumuladas das empresas presentes no conjunto validação. O modelo com restrição apresentou uma métrica de KS um pouco menor, de 39,9%.

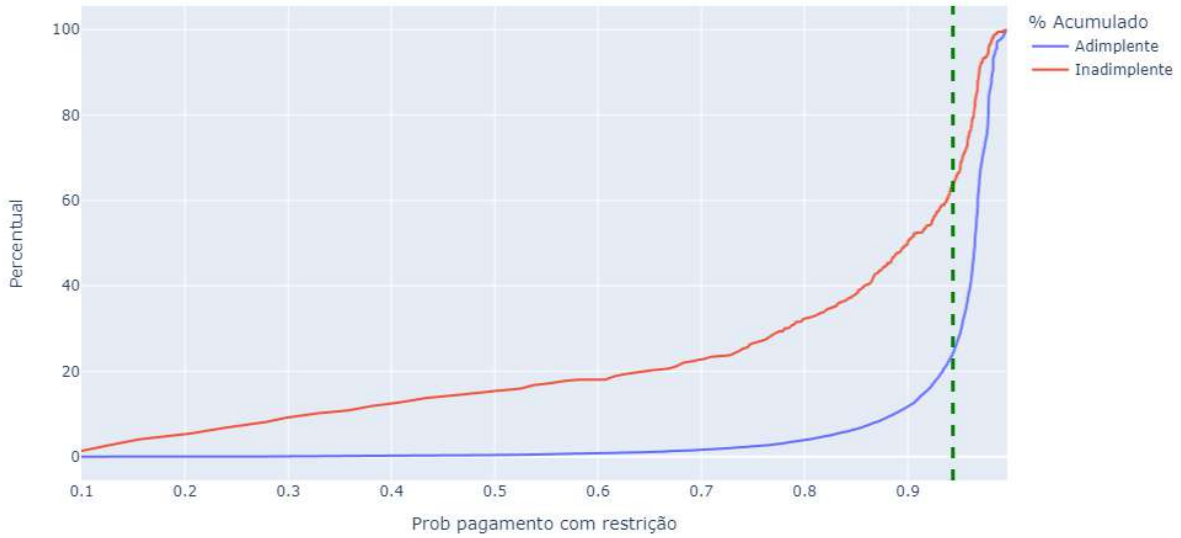

Figura 18 – Distribuição acumulada pela Probabilidade do Modelo com restrição

Comparando as figuras 19 e 12 percebemos que os percentuais de inadimplência nas faixas obtidas são muito similares.

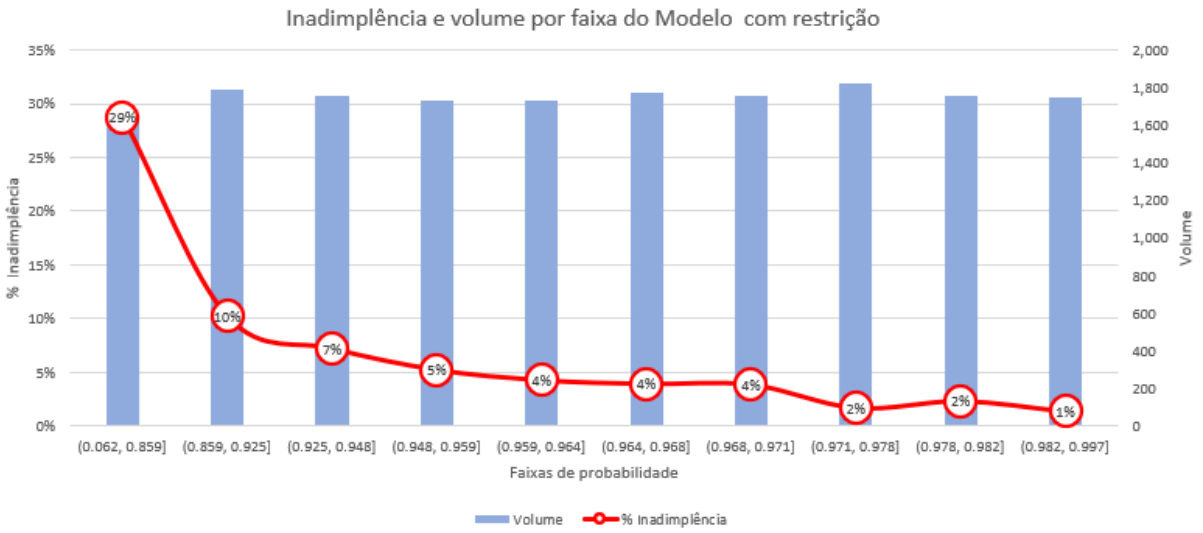

Figura 19 – Gráfico de importância

Para analisar quais variáveis foram mais relevantes para esse novo modelo foi criado o gráfico de importância da imagem 20. É importante observar que a variável sob restrição (IDADE_MENOS_6M) deixou de apresentar importância. Ou seja, a informação contida nela atuava apenas em um sentido. Já a variável IDADE_6Ma1A passou a ser mais importante nesse modelo.

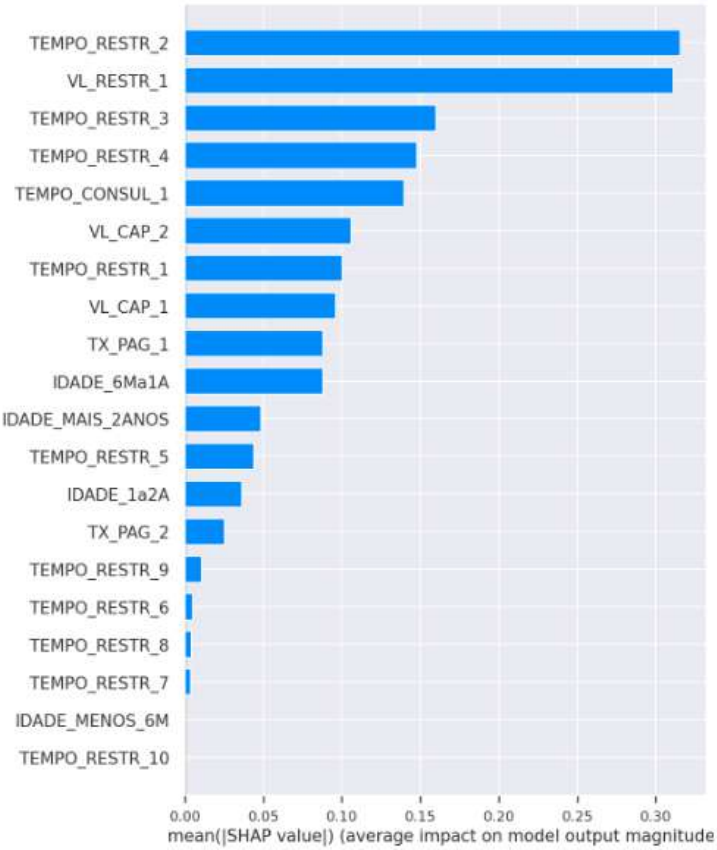

Figura 20 – Gráfico de importância - Modelo com restrição

No gráfico de resumo da figura 21 verificamos em quais sentidos que as 10 variáveis mais importantes nesse novo modelo atuam. Como pode ser notado, a variável IDADE_6Ma1A atua em um sentido diferente do que a IDADE_MENOS_6M no modelo sem restrição. Ou seja, o fato de uma empresa estar na faixa entre 6 meses e 1 ano de fundação, nesse novo modelo, contribui para que ela seja mais arriscada que as demais enquanto que empresas ainda mais novas (menos do que 6 meses) não têm impacto devido a essa característica.

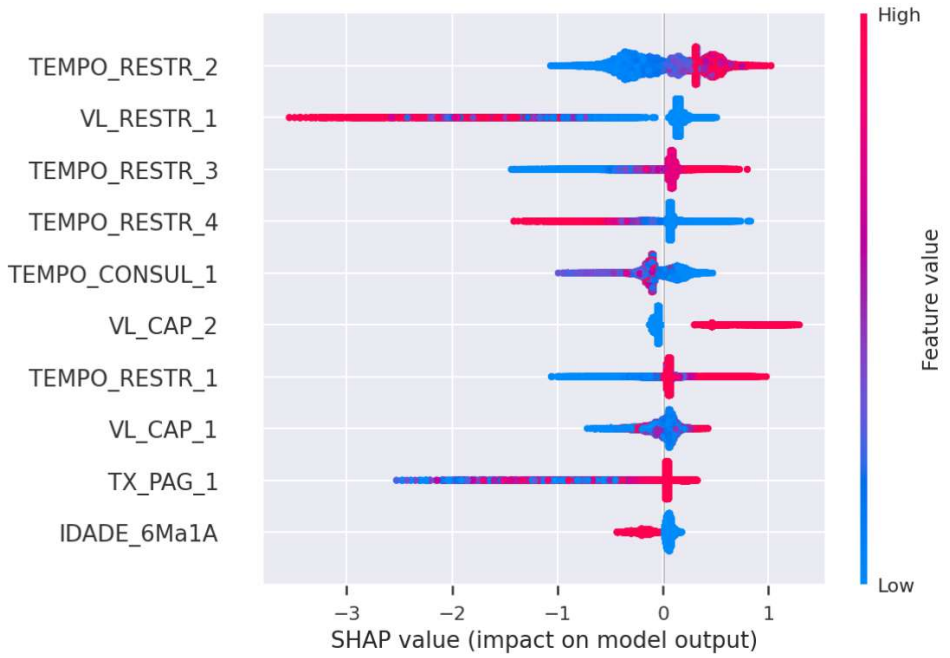

Figura 21 – Gráfico de resumo - Modelo com restrição

Nos gráficos de *boxplot* da figura 22 verificamos as distribuições das probabilidades nesse modelo com restrição, de acordo com a faixa de idade das empresas. Mesmo sem a participação da variável IDADE_MENOS_6M no modelo, empresas que são bem mais recentes têm uma maior concentração em probabilidades que indicam menor risco de concessão de crédito, um comportamento muito similar ao modelo que não tinha nenhuma restrição.

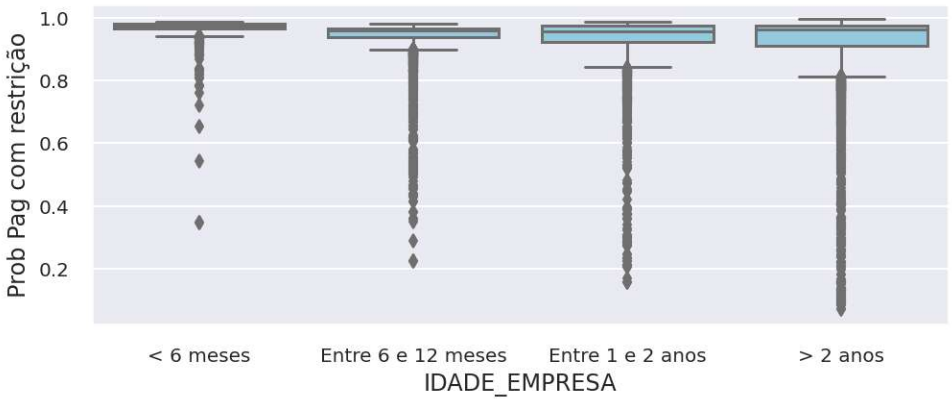

Figura 22 – Boxplot das probabilidades por idade - Modelo com restrição

Na figura 23 foi obtido o gráfico de força para a mesma empresa que foi usada como exemplo no gráfico de força da figura 17. É interessante observar que a probabilidade estimada em ambos os modelos foi a mesma (95%) e que nesse caso a variável indicadora IDADE_6Ma1A ser 0 que impactou positivamente.

Figura 23 – Gráfico de força - Modelo com restrição

Analisando o gráfico de força da figura 23 e as distribuições das probabilidades do modelo com restrição indicada na figura 22, é possível notar que mesmo sem utilizar a informação se a empresa possui ou não menos de 6 meses de fundação o algoritmo combinou os demais dados de forma que empresas que se encaixam nessa faixa de idade tenham uma distribuição de probabilidade distinta de empresas mais velhas.

6.7 Conclusão

Em muitas circunstâncias, o conjunto de dados disponível para o ajuste de um modelo não necessariamente vai apresentar informações totalmente válidas para o contexto em que se deseja gerar valor a partir de suas previsões.

No exemplo da cooperativa de crédito, muito provavelmente foi adotado alguma política em que empresas mais novas são selecionadas de uma forma mais criteriosa, o que resultou em um risco observado menor para esse grupo. É preciso avaliar muito bem se esse conjunto de dados é representativo em relação às novas concessões. Caso a política de crédito seja alterada e novas decisões de concessão sejam feitas baseadas em um modelo desenvolvido com dados da política anterior, é muito provável que os níveis de inadimplência sejam bem diferentes do esperado.

Foi apresentada uma forma de intervir na construção do modelo, restringindo a maneira que o algoritmo seleciona e utiliza as informações das variáveis preditoras. Porém, como o modelo é uma generalização do que é observado no conjunto de dados, mesmo aplicando essas restrições o algoritmo pode encontrar relações entre as demais informações e ajustar um modelo que apresenta um comportamento similar ao que foi construído sem restrição. É uma situação em que ocorre um super ajustamento aos dados, mas muito mais difícil de identificar.

No modelo desenvolvido para a cooperativa de crédito, foi identificado uma relação entre idade das empresas e risco de inadimplência diferente do esperado. A comparação com um *score* genérico confirmou que o risco estimado para empresas com menos de 6 meses de fundação tinham distribuições de risco diferentes nos dois modelos. No *score* genérico essas empresas

eram as mais arriscadas em relação às demais, enquanto o modelo desenvolvido as colocavam como menos arriscadas. Neste contexto, a utilização de metodologias de interpretabilidade, como a dos valores *SHAP*, viabilizam visualizar o comportamento das variáveis, identificar vieses nos dados e discutir quais são as limitações do modelo. Neste caso, a cooperativa de crédito deve manter critérios adicionais além do *score*, principalmente para empresas mais novas, para evitar a concessão de crédito com um risco subestimado.

CAPÍTULO

7

ESPECTROSCOPIA

Neste capítulo, apresentaremos um estudo de caso em que a utilização de metodologias de interpretabilidade de modelos complexos pode auxiliar pesquisadores a investigar fenômenos de uma forma mais detalhada. Neste exemplo serão usados dados de espectroscopia para estimar a concentração de carbono em amostras de solo.

7.1 Dados

O presente estudo empregou dados obtidos com a técnica de *LIBS* (*laser induced breakdown spectroscopy*). Esta técnica fornece informações sobre a intensidade da radiação emitida após a aplicação de pulsos de laser no solo. Um total de 1.019 amostras de solo foi coletado de profundidades variando entre 0 e 100 cm, oriundas de onze propriedades brasileiras. Estas propriedades abrangem vegetações nativas e solos agrícolas de três biomas distintos: cerrado, floresta atlântica e pampa. Posteriormente, as amostras foram submetidas a um processo de secagem, homogeneização e prensagem em pastilhas, preparando-as para a análise pelo sistema LIBS. Uma fração destas amostras secas e homogeneizadas foi destinada à quantificação de carbono, utilizando a técnica de combustão a seco com o equipamento CHNS da Perkin Elmer.

Para cada amostra, 30 medições LIBS foram realizadas em pontos diferentes, resultando na coleta de um espectro por ponto. O equipamento utilizado para estas medições possui um laser Nd:YAG de 1064 nm e um conjunto de sete espectrômetros HR2000+ da Ocean Optics. No desenvolvimento dos modelos de aprendizado de máquina, a média dos espectros de cada amostra foi usada. Esta média representa a intensidade média para cada comprimento de onda registrado pelos espectrômetros (13.748 no total). Assim, o conjunto de dados é composto pelo espectro médio e pela concentração de carbono de cada amostra, totalizando 13.749 variáveis. Um exemplo de espectro é apresentado na figura 24.

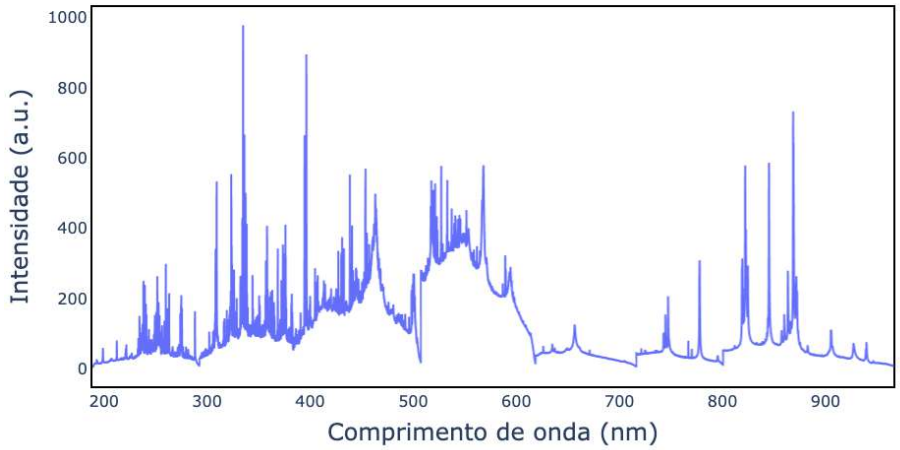

Figura 24 – Exemplo de espectro médio obtido de uma amostra

7.2 Objetivo do modelo

A partir dos espectros presentes nesses dados, é possível estimar a concentração de carbono presente na amostra. Isso é vantajoso porque se torna uma forma rápida e barata de identificar esse elemento no solo (EBINGER *et al.*, 2003). No entanto, é preciso desenvolver modelos que relacionem os dados que foram extraídos de maneira eficiente. Outro ponto bastante relevante é que, além de avaliar a concentração de carbono em uma amostra, é possível examinar as informações que conduziram o modelo a prever essa concentração. Por exemplo, solos com maior teor de areia tendem a facilitar a decomposição de matéria orgânica, resultando em menor concentração de carbono. Em tal cenário, abordagens de interpretabilidade de modelos podem revelar, além da característica da concentração de carbono, outras informações importantes sobre a composição do solo.

7.3 Seleção de variáveis

Usualmente, o volume de informação contido em dados de espectroscopia é bastante elevado. O espectro de emissão gerado em cada amostra a ser analisada contém dados de uma ampla faixa de comprimentos de onda, o que leva à criação de conjuntos de dados com milhares de variáveis candidatas a participar no modelo.

Diante dessa situação, a seleção de variáveis torna-se uma etapa importante na modelagem. Uma metodologia amplamente utilizada é conhecida como *"wrapper"*, na qual um modelo preliminar é ajustado usando um algoritmo de aprendizado de máquina, levando em consideração todas as variáveis disponíveis. Ao contrário do modelo final, o objetivo desse modelo preliminar é simplesmente identificar as variáveis mais relevantes para descrever a variável resposta (JOVIĆ; BRKIĆ; BOGUNOVIĆ, 2015).

Uma das abordagens de pré-seleção por *wrapper* é baseada no algoritmo de florestas aleatórias, no qual são empregadas diversas árvores de decisão por meio de um *ensemble* para construir uma predição final (MUSTAQEEM *et al.*, 2017). Neste trabalho foi empregado um modelo prévio de floresta aleatória considerando amostragem das variáveis disponíveis, com reposição, na construção das árvores para identificar aquelas que mais participaram na discriminação desse modelo, como forma de separação entre as variáveis mais e menos relevantes. Através dele foi possível selecionar nos dados *LIBS* 1,098 variáveis das 13,748 disponíveis. Esta etapa permite que a demanda de recursos computacionais no desenvolvimento do modelo seja consideravelmente reduzida, além de direcionar quais informações são mais relevantes e auxiliar na construção de um modelo mais generalizável.

7.4 Ajuste do modelo

Neste caso a variável resposta é contínua, com isso o modelo a ser desenvolvido é de regressão. O algoritmo utilizado foi a implementação de modelos de regressão do *Lightgbm*. O conjunto de dados foi dividido em desenvolvimento e validação, na proporção 70 - 30%, respectivamente.

A partir das variáveis selecionadas foi feita uma busca dos hiperparâmetros do modelo através do algoritmo *optuna*. O modelo final desenvolvido apresentou os seguintes valores de hiperparâmetros:

Hiperparâmetro	Valor
n_estimator	194
learning_rate	0.05256
max_depth	32
num_leaves	102

Tabela 4 – Parâmetros do modelo de predição de concentração de carbono.

O hiperparâmetro *n_estimator* aponta que esse modelo contém 194 árvores, o *max_depth* limita a profundidade dessas árvores a 32 níveis e o num_leaves de 102 o número de folhas. Esses números evidenciam a complexidade associada a esse modelo de *boosting*.

7.5 Resultados

As métricas selecionadas para apontar o quanto o modelo se ajustou aos dados foram: MAE (erro médio absoluto), RMSE (raiz do erro quadrático médio) e R^2 (Coeficiente de determinação). Na tabela 5 são apresentadas essas métricas no conjunto de validação.

Métrica	Valor obtido na validação
MAE	3,1514
RMSE	4,5412
R^2	0,7686

Tabela 5 – Métricas do modelo

Na figura 25 é possível visualizar a dispersão das predições tanto no conjunto de ajuste (desenvolvimento do modelo) quanto de avaliação de acordo com a concentração de carbono observada na amostra. Quanto mais próximo o ponto da linha diagonal tracejada, mais precisa foi a predição do modelo em relação a concentração medida através do método CHNS.

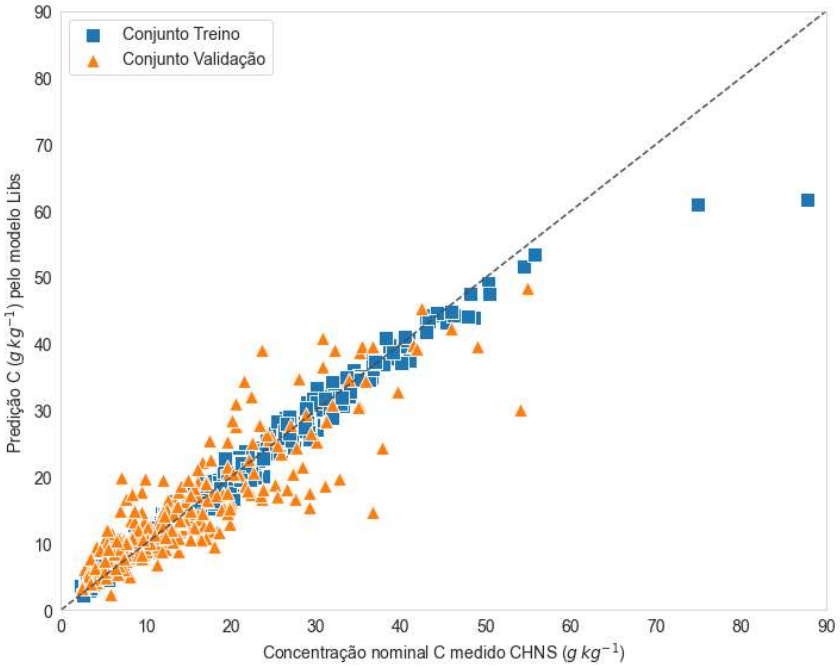

Figura 25 – Gráfico de dispersão

7.6 Interpretação

Com o desenvolvimento do modelo, foi feita uma análise dos valores shap das linhas de emissão mais importantes e identificados quais são os seu elementos associados a partir do banco de dados atômicos espectrais NIST (KRAMIDA *et al.*, 2022). Dessa forma, os gráficos já foram gerados com o nome do elemento químico associado à linha de emissão. O gráfico de importância da figura 26 aponta as 20 linhas de emissão mais importantes para o modelo desenvolvido. Nota-se que as linhas associadas ao próprio carbono (C I) 193.04 e 192.98 foram as que mais se destacaram. É interessante comentar que outros estudos também relacionam essas linhas ao carbono (EBINGER *et al.*, 2003). Além disso, observa-se que a partir da décima linha destacada, a 634.63 (Si II), não há grande diferença de importância entre as variáveis.

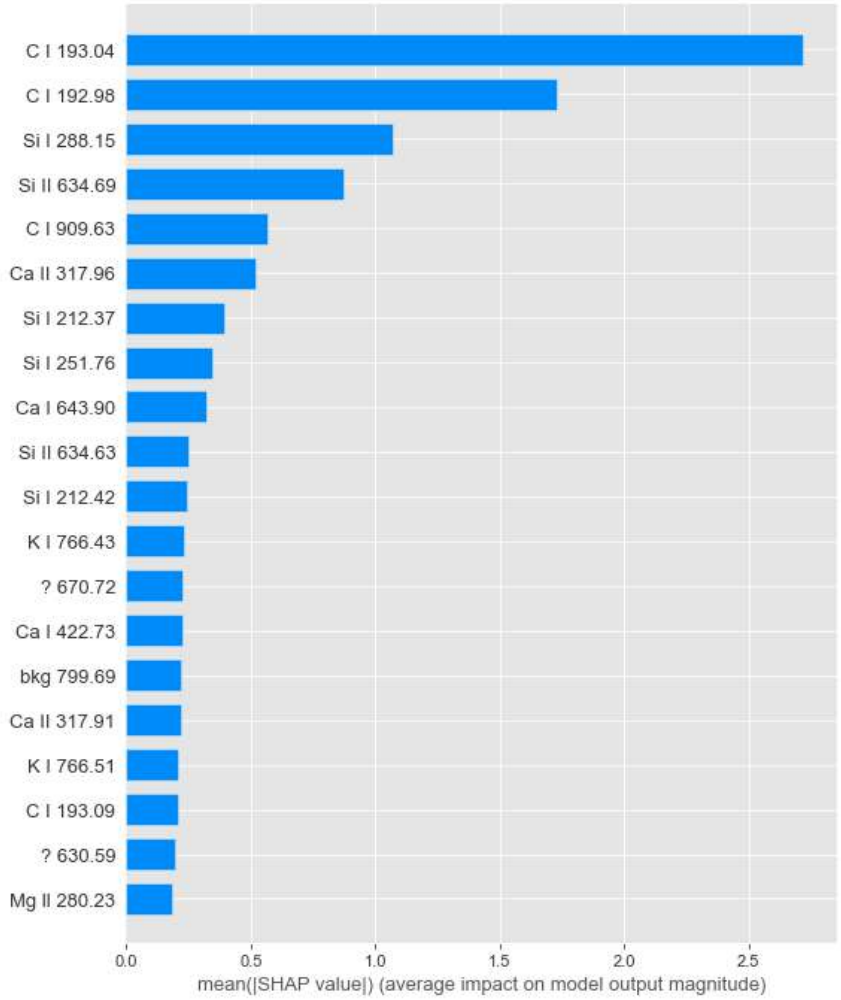

Figura 26 – Gráfico de importância

Na figura 27 é apresentado o gráfico de resumo na amostra de validação para o modelo de predição de carbono. A escala de cor indica se o valor da variável (linha de emissão) observada é alto ou baixo. A linha de emissão mais relevante é a 193.04 nm, do próprio carbono (C I). As cores indicam que valores mais altos nessa linha de emissão (os pontos vermelhos), na maior parte dos casos impacta no cálculo do modelo de forma a aumentar a predição de carbono, enquanto valores mais baixos (pontos azuis) levam a redução da estimação. Esse comportamento é similar para as linhas 192.98 (C I), 909.63 (C I), 670.72 (?), 422.73 (Ca I) e 193,09 (C I). Para outras linhas de emissão o comportamento é contrário ao mencionado, isto é, quanto mais elevado o valor observado na linha de emissão o impacto dessa informação no modelo leva a estimação de uma menor concentração de carbono. As linhas que apresentaram esse tipo de comportamento no modelo foram: 288.15 (Si I), 634.69 (Si II), 212.37 (Si I), 634.63 (Si II), 212.42 (Si I), 799.69 (bkg) e 630.59 (?). Há ainda alguns casos que não têm a separação entre as cores muito clara, o que indica que, para o modelo, a contribuição da informação que foi obtida nessa linha de emissão depende da combinação com as demais características da amostra, como no caso das linhas 317.91 (Ca I) e 280.23 (Mg II).

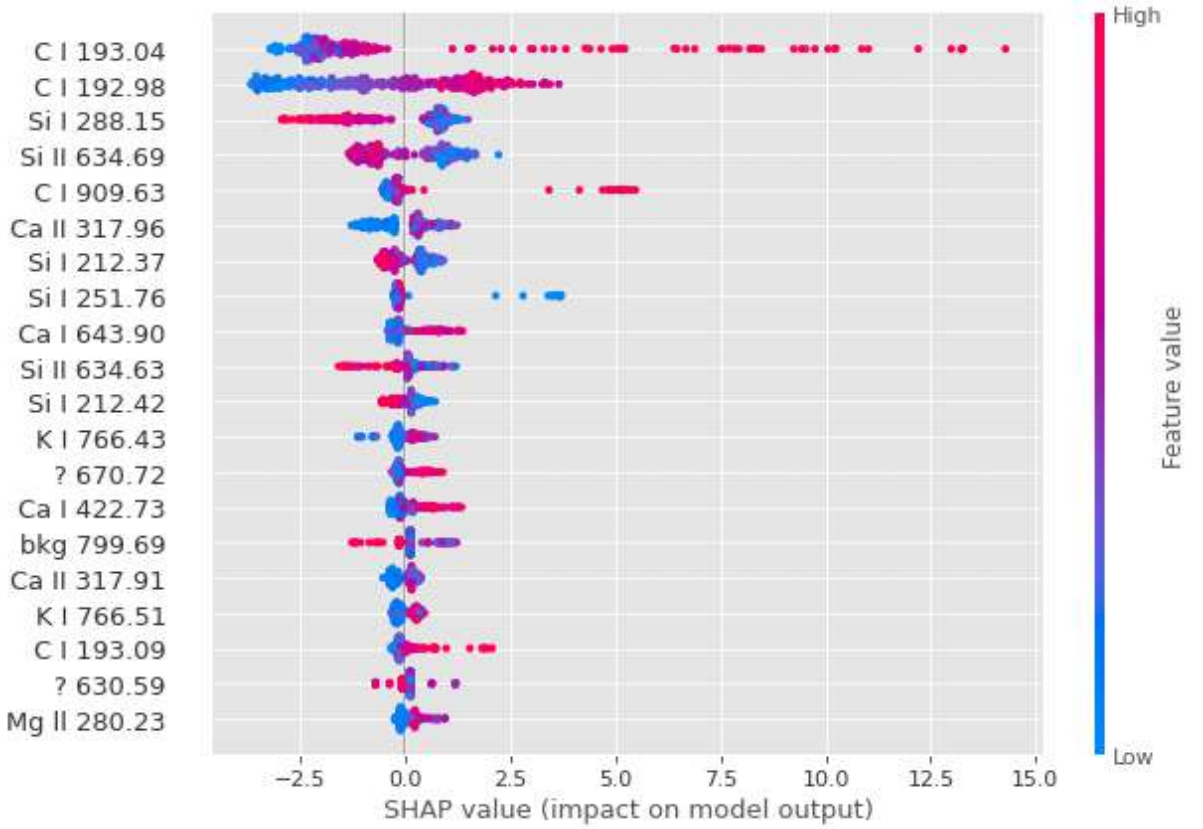

Figura 27 – Gráfico de resumo

Foram separados também 4 exemplos de amostras para indicar quais foram as linhas de emissão mais relevantes em cada caso através do gráfico de força. Como não foi feito nenhum pré processamento nesse conjunto de dados para desenvolvimento do modelo é preciso ter uma referência de qual é a faixa de variação em cada linha de emissão. O exemplo de espectro apresentado na figura 24 mostra o quanto a intensidade em cada comprimento de onda pode ser diferente. Para facilitar a avaliação se um determinado valor é alto ou baixo em uma determinada variável nos gráficos de força, foram apontados os percentis que os valores da variáveis se encontram nesse conjunto de dados.

Na figura 28 são indicados os gráficos de força. São 4 amostras, todas do conjunto de validação e que apresentaram quantidade de carbono por kg muito diferentes. Como pode ser visto, o *base value* desse modelo é 15.43g kg^{-1}, o que indica o valor médio de concentração de carbono que foi observado no treino desse modelo. A amostra *a* apresentou concentração medida pelo método CHNS de 3g kg^{-1} (essa informação não é revelada no gráfico de força), enquanto que o modelo apontou 4.34g kg^{-1}. Nota-se que praticamente nenhuma variável contribuiu significativamente para que o valor predito do modelo fosse mais alto, no gráfico de força nenhuma linha de emissão é destacada na parte em vermelho. Já no sentido de reduzir a predição da concentração de carbono participaram as linhas associadas ao próprio carbono (C I), 193.04179 e 192.98664, com valores baixos (percentis 19 e 21, respectivamente). Além disso, também contribuíram para reduzir a predição as linhas 288.15793 (Si I) e 634.69 (Si II), com valores altos

(P94 e P81). Além de observar baixa concentração de carbono nesta amostra, há forte evidência de presença de silício.

Já na amostra *b* da figura 28, foi medida uma concentração de 10g kg^{-1} pelo método CHNS e apontado 9.84g kg^{-1} pelo modelo. Em comparação com a amostra *a*, observamos que as linhas 634.69 (Si II) e 288.15 (Si I) apresentaram valores menores (P55 e P61). Esses números sugerem que essa amostra apresenta uma quantidade menor de silício do que a amostra *a*, o que levou o modelo a predizer uma concentração de carbono um pouco maior.

O gráfico de força *c* demonstra a estimação de 19.6g kg^{-1} feita pelo modelo em uma amostra que foi medida uma concentração de 15 kg^{-1}. As linhas de emissão do C I, 193.04 e 192.98, apresentaram valores altos, assim como a linha 422.73 (Ca II), enquanto que a linha do silício 288.15 (Si I) apresentou um valor próximo da mediana observada (percentil 52). Todas essas informações contribuíram para aumentar a estimação de carbono. No sentido de reduzir essa estimação destaca-se a linha 766.43 (K I), com um valor baixo, percentil 24.

Por último, para a amostra *d* o modelo estimou uma concentração de carbono de 39.29g kg^{-1}, sendo que foi medido pelo método CNHS 41.8g kg^{-1}. Neste caso, observa-se valores muito altos nas linhas do carbono, 193.04 e 192.98, assim como um valor abaixo da mediana na linha 799.69 (bkg). Nenhuma linha de emissão no sentido de reduzir a estimação (parte azul) é destacada para esta amostra.

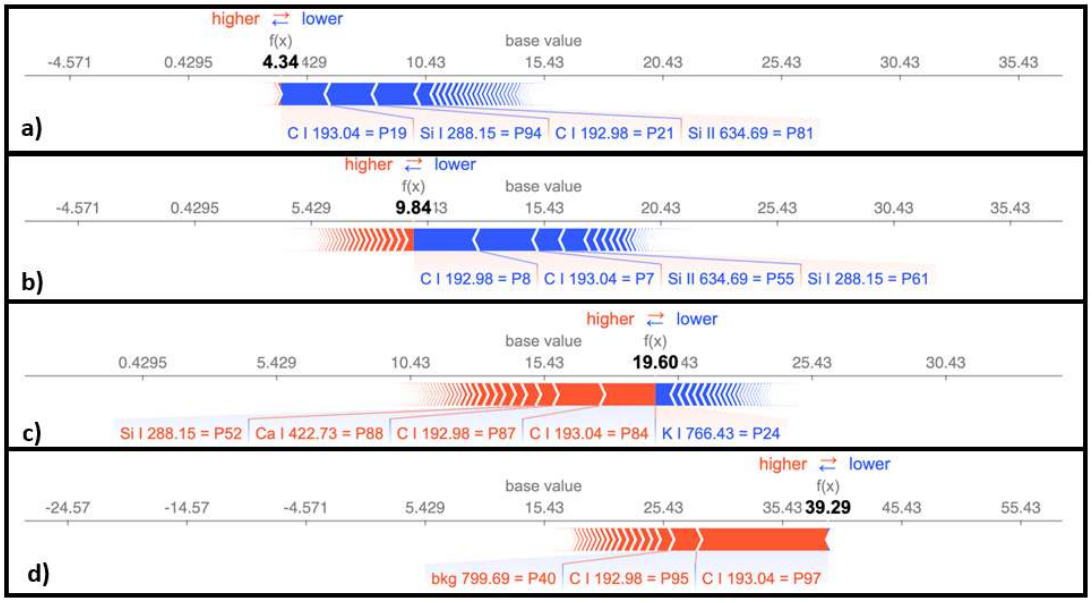

Figura 28 – Gráficos de força de 4 amostras distintas

7.7 Conclusão

Modelos complexos de aprendizado de máquina podem ser utilizados para a identificação de elementos químicos em amostras de solo com base em dados de espectroscopia. No estudo

de caso apresentado, foi desenvolvido um modelo de *boosting* para estimar a concentração de carbono em amostras de solo coletadas em diferentes biomas brasileiros. A interpretação desses modelos pode fornecer informações adicionais sobre a composição desses solos.

Ao examinar os gráficos de força, identificamos evidências da presença ou ausência de outros elementos químicos, como silício, cálcio e potássio em diferentes amostras de solo. É importante destacar que não foi necessário realizar ajustes ou pré-processamento nos dados, o que acelera significativamente o processo de identificação dos elementos e a obtenção de informações sobre a composição dos solos.

CAPÍTULO

8

CONCLUSÃO

Metodologias que buscam interpretabilidade de modelos complexos de aprendizado de máquina são muito importantes para viabilizar uma melhor compreensão sobre a forma que as informações contidas nos dados estão sendo relacionadas. Um modelo é uma generalização construída a partir dos dados que foram apresentados no seu treinamento, se não for feita uma avaliação dessa generalização, isto é, como que as características impactam o funcionamento do modelo, há um grande risco desses modelos continuarem propagando padrões que não são válidos e/ou se tornaram obsoletos. Essa generalização pode ser útil de diversas formas, como na busca de uma melhor compreensão sobre os eventos em questão (inferência) ou identificação de novos eventos (predição), e utilizar métodos de interpretação também pode ser proveitoso de várias maneiras.

Neste trabalho foram utilizados os valores *SHAP* para estimar e visualizar os impactos das variáveis preditoras em modelos de diferentes contextos. No caso do Exame Nacional de Ensino Médio, foi interessante identificar quais são as características mais relevantes, de acordo com o que foi aprendido pelo modelo, para separar candidatos mais ou menos propensos a atingirem uma nota mínima no exame. Além disso, foram levantados alguns exemplos de candidatos para poder interpretar como suas características socio econômicas eram utilizadas na generalização construída pelo modelo de aprendizado de máquina. Neste exemplo, buscou-se compreender melhor o conjunto de dados e avaliar quais são os padrões identificados. Fatores como renda familiar, escolaridade dos pais, acesso a computador, cor de pele e sexo influenciam na probabilidade do candidato obter uma nota mínima no exame.

Para o caso de concessão de crédito o objetivo principal era avaliar se o que foi aprendido pelo modelo estava dentro do que é esperado. Como o objetivo desse modelo é predição, com o propósito de antecipar quais empresas têm um maior risco de não pagar o valor que foi concedido como empréstimo, é muito importante analisar se a generalização construída pelo modelo é aplicável para todo o mercado. Diferentemente do que normalmente é visto no mercado, o modelo

que foi desenvolvido com os dados da cooperativa de crédito indicava que empresas muito novas eram menos arriscadas. Muito provavelmente esse tipo de relação é devido a algum viés da amostra que foi disponibilizada, ou seja, provavelmente as empresas que obtiveram crédito e que tinham essas características foram pré selecionadas em alguma política de crédito mais criteriosa, o que refletiu em taxas de inadimplência que não são representativas para esse grupo. Além disso, houve uma tentativa de criar um modelo com restrição de monotonicidade para essa característica, porém esse novo modelo se ajustou aos dados a partir de outras variáveis e também indicava riscos menores para empresas muito novas. Em situações como essa, interpretar o modelo é muito importante para avaliar suas limitações e adequar o seu uso.

Por último, foi apresentada uma aplicação de modelagem em dados de espectroscopia como uma forma de identificar a concentração de carbono em amostras de solo. Neste contexto, os valores *SHAP* foram calculados para trazer informações adicionais às reveladas diretamente pelo modelo, como quais linhas de emissão foram mais relevantes e a forma que elas contribuíram nas estimações da concentração de carbono. Como essas linhas são associadas a elementos químicos específicos, através da interpretação desse modelo é possível ter uma melhor compreensão sobre a composição do solo.

Por meio desses exemplos, torna-se evidente que é possível analisar os padrões que foram aprendidos por modelos complexos de forma mais detalhada, identificar vieses presentes no conjunto de dados, compreender as limitações desses modelos e até mesmo interferir no processo de aprendizado para que eles sejam mais performáticos (reduzindo super ajustamento, por exemplo). Assim, metodologias de interpretação como a dos valores *SHAP* podem melhorar a transparência e a confiabilidade dos modelos de aprendizado de máquina, além de ajudar os usuários a entender melhor como o modelo está tomando suas decisões.

REFERÊNCIAS

ALMEIDA, V. S. de. O encceja e o enem: o exame nacional do ensino médio como ferramenta para certificação do ensino médio. **Saberes Interdisciplinares**, v. 13, n. 25, p. 11–22, 2020. Citado na página 42.

ATHEY, S. The impact of machine learning on economics. In: **The economics of artificial intelligence: An agenda**. [S.l.]: University of Chicago Press, 2018. p. 507–547. Citado nas páginas 25 e 26.

BENTO, M. A. S. Branquitude e poder-a questão das cotas para negros. 2005. Citado na página 47.

BRYNJOLFSSON, E.; MITCHELL, T. What can machine learning do? workforce implications. **Science**, American Association for the Advancement of Science, v. 358, n. 6370, p. 1530–1534, 2017. Citado na página 21.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: **Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining**. [S.l.: s.n.], 2016. p. 785–794. Citado na página 27.

COGLIANESE, C.; LEHR, D. Regulating by robot: Administrative decision making in the machine-learning era. **Geo. LJ**, HeinOnline, v. 105, p. 1147, 2016. Citado nas páginas 21, 26 e 27.

DOSHI-VELEZ, F.; KIM, B. Towards a rigorous science of interpretable machine learning. **arXiv preprint arXiv:1702.08608**, 2017. Citado na página 21.

EBINGER, M. H.; NORFLEET, M. L.; BRESHEARS, D. D.; CREMERS, D. A.; FERRIS, M. J.; UNKEFER, P. J.; LAMB, M. S.; GODDARD, K. L.; MEYER, C. W. Extending the applicability of laser-induced breakdown spectroscopy for total soil carbon measurement. **Soil Science Society of America Journal**, Wiley Online Library, v. 67, n. 5, p. 1616–1619, 2003. Citado nas páginas 62 e 64.

FORTI, M. **Técnicas de machine learning aplicadas na recuperação de crédito do mercado brasileiro**. Tese (Doutorado), 2018. Citado nas páginas 22 e 49.

GÉRON, A. Hands-on machine learning with scikit-learn and tensorflow: Concepts. **Tools, and Techniques to build intelligent systems**, O'Reilly Media, 2017. Citado nas páginas 26 e 27.

HARARI, Y. N. **21 lições para o século 21**. [S.l.]: Editora Companhia das Letras, 2018. Citado nas páginas 22 e 23.

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Microdados do Exame Nacional do Ensino Médio - ENEM**. 2021. Disponível em: <https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem>. Citado na página 41.

IZBICKI, R.; SANTOS, T. M. dos. **Aprendizado de máquina: uma abordagem estatística**. [S.l.]: Rafael Izbicki, 2020. Citado nas páginas 26 e 27.

JOVIĆ, A.; BRKIĆ, K.; BOGUNOVIĆ, N. A review of feature selection methods with applications. In: IEEE. **2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO)**. [S.l.], 2015. p. 1200–1205. Citado na página 62.

KIM, S. H.; BOUKOUVALA, F. Machine learning-based surrogate modeling for data-driven optimization: a comparison of subset selection for regression techniques. **Optimization Letters**, Springer, p. 1–22, 2019. Citado na página 30.

KRAMIDA, A.; RALCHENKO, Y.; READER, J.; NIST ASD Team. **NIST Atomic Spectra Database (version 5.10), [Online]**. 2022. National Institute of Standards and Technology, Gaithersburg, MD (Available: https://physics.nist.gov/asd [2023, September 07]). Citado na página 64.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **nature**, Nature Publishing Group UK London, v. 521, n. 7553, p. 436–444, 2015. Citado na página 27.

LIPTON, Z. C. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. **Queue**, ACM New York, NY, USA, v. 16, n. 3, p. 31–57, 2018. Citado nas páginas 21 e 29.

LUNDBERG, S.; LEE, S.-I. A unified approach to interpreting model predictions. **arXiv preprint arXiv:1705.07874**, 2017. Citado nas páginas 23, 33, 34 e 35.

LUNDBERG, S. M.; ERION, G. G.; LEE, S.-I. Consistent individualized feature attribution for tree ensembles. **arXiv preprint arXiv:1802.03888**, 2018. Citado nas páginas 34 e 35.

MESSALAS, A.; KANELLOPOULOS, Y.; MAKRIS, C. Model-agnostic interpretability with shapley values. In: IEEE. **2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)**. [S.l.], 2019. p. 1–7. Citado na página 31.

MOLNAR, C. **Interpretable Machine Learning**. 2021. Disponível em: <https://christophm.github.io/interpretable-ml-book/shap.html#treeshap>. Citado nas páginas 15, 35, 36, 37, 38 e 39.

MOLNAR, C.; CASALICCHIO, G.; BISCHL, B. Interpretable machine learning–a brief history, state-of-the-art and challenges. **arXiv preprint arXiv:2010.09337**, 2020. Citado nas páginas 23, 24, 29, 30 e 31.

MUSTAQEEM, A.; ANWAR, S. M.; MAJID, M.; KHAN, A. R. Wrapper method for feature selection to classify cardiac arrhythmia. In: IEEE. **2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)**. [S.l.], 2017. p. 3656–3659. Citado na página 63.

NAQA, I. E.; MURPHY, M. J. **What is machine learning?** [S.l.]: Springer, 2015. Citado na página 25.

OLIVEIRA, L. H. R. de. Mercado de crédito: a importância do mercado de crédito para a economia brasileira. 2018. Citado nas páginas 22 e 49.

PALMA, L. Agrupamento de dados: k-médias. **Universidade Federal do Recôncavo da Bahia Centro de Ciências Exatas e Tecnológicas**, 2018. Citado na página 27.

PICCIN, L. E. Métodos de detecção de fraude em cartões de crédito: um estudo comparativo. Universidade Federal de São Carlos, 2022. Citado na página 50.

PODGORELEC, V.; KOKOL, P.; STIGLIC, B.; ROZMAN, I. Decision trees: an overview and their use in medicine. **Journal of medical systems**, Springer, v. 26, p. 445–463, 2002. Citado na página 26.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. "why should i trust you?"explaining the predictions of any classifier. In: **Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining**. [S.l.: s.n.], 2016. p. 1135–1144. Citado na página 30.

RODRÍGUEZ-PÉREZ, R.; BAJORATH, J. Interpretation of compound activity predictions from complex machine learning models using local approximations and shapley values. **Journal of Medicinal Chemistry**, ACS Publications, v. 63, n. 16, p. 8761–8777, 2019. Citado na página 33.

ROTH, A. E. Lloyd shapley (1923–2016). **Nature**, Nature Publishing Group UK London, v. 532, n. 7598, p. 178–178, 2016. Citado na página 33.

SAKALAUSKAS, S. R.; TREVISAN, A. L. Enem: rompendo paradigmas para a conclusão do ensino médio. **[TESTE] Debates em Educação**, v. 9, n. 19, p. 01, 2017. Citado na página 41.

SAMPAIO, B.; GUIMARÃES, J. Diferenças de eficiência entre ensino público e privado no brasil. **Economia Aplicada**, SciELO Brasil, v. 13, p. 45–68, 2009. Citado na página 47.

SHARMA, A.; WEHRHEIM, H. Higher income, larger loan? monotonicity testing of machine learning models. In: **Proceedings of the 29th ACM SIGSOFT International Symposium on Software Testing and Analysis**. [S.l.: s.n.], 2020. p. 200–210. Citado na página 55.

SILVEIRA, F. L. d.; BARBOSA, M. C. B.; SILVA, R. d. **Exame Nacional do Ensino Médio (ENEM): uma análise crítica**. [S.l.]: SciELO Brasil, 2015. Citado na página 41.

UNCETA, I.; NIN, J.; PUJOL, O. Towards global explanations for credit risk scoring. **arXiv preprint arXiv:1811.07698**, 2018. Citado na página 23.

WEI, P.; LU, Z.; SONG, J. Variable importance analysis: a comprehensive review. **Reliability Engineering & System Safety**, Elsevier, v. 142, p. 399–432, 2015. Citado na página 31.


ICMC
USP
SÃO CARLOS