Estimando a Arrecadação da Dívida Ativa da União com Machine Learning: Uma análise baseada nos dados de arrecadação do período de 2015 a 2021[1]

Estimating the Union's Active Debt Revenues with Machine Learning*: An analysis based on tax revenues data from 2015 to 2021.*

Estimación de los ingresos de la deuda activa de la Unión con Machine Learning*: un análisis basado en datos de ingresos fiscales de 2015 a 2021.*

Rubens Quaresma Santos

https://doi.org/10.36428/revistadacgu.v14i26.529

Resumo: Como entidade responsável pela cobrança da Dívida Ativa da União, a Procuradoria-Geral da Fazenda Nacional deve apresentar, ao final de cada exercício fiscal, os resultados alcançados e as previsões de arrecadação para o ano seguinte para a composição das Leis Orçamentárias vindouras. Atualmente essa estimativa é elaborada utilizando a técnica da suavização exponencial, que considera os ingressos passados para projetar a arrecadação futura. O presente trabalho procura avaliar a aplicação de algoritmos de aprendizado de máquina na elaboração dessas projeções, como meio de aprimoramento da gestão. Foram testados os algoritmos Regressão Linear, Árvore de Decisão, Floresta Randômica e Árvore de Decisão de Aumento de Gradiente. Os modelos foram alimentados com informações dos indicadores macroeconômicos IPCA, IGP-M, PIB, taxa de Câmbio e taxa SELIC, além de dados relativos a parcelamentos excepcionais e transações tributárias disponibilizados aos contribuintes pela Fazenda Nacional.

Palavras-chave: Dívida Ativa. Arrecadação Fiscal. Orçamento. Machine Learning. Ciência de Dados.

Abstract: As the entity responsible for collecting the Active Debt of the Union, the Attorney General's Office of the National Treasury must present, at the end of each fiscal year, the results achieved and the collection forecasts for the following year for the composition of forthcoming Budget Laws. Currently, this estimate is made using the exponential smoothing technique, which considers past income to project future revenue. This article seeks to evaluate the application of machine learning algorithms in the elaboration of these projections, as a means of improving management. Linear Regression, Decision Tree, Random Forest and Gradient Boosting Decision Tree algorithms were tested. The models were fed with information from the macroeconomic indicators IPCA, IGP-M, GDP, Exchange rate and SELIC rate, as well as data on exceptional installments and tax transactions made available to taxpayers by the National Treasury.

Keywords: Active debt. Tax Collection. Budget. Machine Learning. Data Science.

Resumen: La Procuraduría General del Tesoro Nacional, cómo ente encargado de recaudar la Deuda Activa de la Unión debe presentar, al cierre de cada ejercicio fiscal, los resultados alcanzados y las previsiones de recaudación del ejercicio siguiente para la composición de las próximas Leyes de Presupuesto. Actualmente, esta estimación se

1. Artigo submetido em 11/07/2022 e aceito em 01/12/2022.

realiza mediante la técnica de suavizamiento exponencial, que considera los ingresos pasados para proyectar los ingresos futuros. Este artículo busca evaluar la aplicación de algoritmos de aprendizaje automático en la elaboración de estas proyecciones, cómo medio para mejorar la gestión. Se probaron los algoritmos de regresión lineal, árbol de decisión, bosque aleatorio y árbol de decisión de aumento de gradiente. Los modelos fueron alimentados con información de los indicadores macroeconómicos IPCA, IGP-M, PBI, Tipo de Cambio y Tasa SELIC, así como datos de cuotas excepcionales y operaciones tributarias puestas a disposición de los contribuyentes por parte del Tesoro Nacional.

Palabras clave: Deuda activa. Recaudación de impuestos. Presupuesto. Aprendizaje automático. Ciencia de los datos.

1. INTRODUÇÃO

Na organização administrativa do Estado brasileiro, a Procuradoria-Geral da Fazenda Nacional (PGFN) é a instituição competente para representar a União na execução da sua dívida ativa, conforme expressa dicção do §3º, do art. 131, da Constituição Federal de 1988.

Sendo o Órgão incumbindo da apuração da liquidez, certeza e exigibilidade dos créditos encaminhados para inscrição no passível recebível da União, bem como sua respectiva cobrança, a PGFN é chamada, ao final de cada exercício financeiro, a estimar a arrecadação dos valores da Dívida Ativa da União - DAU para o ano seguinte.

Essa previsão é de suma importância para a composição da Lei Orçamentária Anual e, consequentemente, para a formulação de planos e projetos do Estado Brasileiro. Logo, a precisão e certeza dessa previsão é fundamental, uma vez que é através dela e dos prognósticos de outras fontes de receita que o Poder Executivo Federal elaborará o orçamento governamental e, em última análise, balizará as verbas a serem destinadas às diversas políticas públicas em curso.

Como é possível averiguar da Nota SEI nº 29/2021/COAGED/CDA/PGDAU/PGFN-ME[2], os cálculos atuais tomam por base os valores arrecadados no ano anterior e utilizam a metodologia de suavização exponencial para projetar as entradas para o ano seguinte. Este artigo procura avaliar se é possível, utilizando modelos de Machine Learning, aprimorar essas estimativas arrecadatórias.

As técnicas de aprendizado de máquina têm se mostrado eficientes em avaliar quantidades significativas de dados e encontrar padrões que podem ser utilizados para elaboração de projeções e previsões em diversas áreas, não existindo, a priori, qualquer razão que justifique a Administração Pública não se valer dessas ferramentas para aprimorar sua gestão.

1.1. Suposições e Pressupostos

A premissa deste estudo é que a consideração de variáveis macroeconômicas pode servir de sinalizador dos montantes arrecadados pela DAU, porquanto elas registram características gerais da situação financeira dos agentes econômicos nacionais e a tributação significa, essencialmente, a capitação de recursos por meio da incidência de exações sobre a atividade produtiva.

Por essa razão, na implementação dessa avaliação são considerados o Índice Nacional de Preços ao Consumidor - IPCA, elaborado pelo Instituto Brasileiro de Geografia e Estatística - IBGE; o Índice Geral de Preços do Mercado - IGP-M, apurado pela Fundação Getúlio Vargas - FGV; a flutuação cambial do Dólar; a taxa SELIC; e a variação do Produto Interno Bruto - PIB.

O IPCA registra a inflação da nossa moeda, é calculado com base na variação de preços de um conjunto de produtos e serviços comercializados no varejo e é um importante indicador do custo do capital em nossa sociedade (INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA, 2022a). O IGP-M, ao seu turno, acompanha a variação de preços no mercado e toma em consideração o valor de matérias primas agrícolas e industriais, constitui um indicativo da atividade econômica do país (FUNDAÇÃO GETÚLIO VARGAS, 2022).

A consideração da variação cambial é relevante por registrar um comparativo do valor de nossa moeda em face do Dólar, que a partir da conferência de Bretton Woods em 1944 passou a balizar as transações comerciais internacionais (KRUGMAN E OBSTFELD, 2005,

2. Documento obtido via da Lei de Acesso à Informação, pedido nº. 03005.291447/2022-40.

p. 408). Dado que parte significativa dos produtos e serviços produzidos e consumidos no Brasil são negociados com parceiros estrangeiros (BANCO CENTRAL DO BRASIL, 2022a), sua ponderação representa a consideração dos reflexos do comércio internacional nos custos produtivos e na inflação nacional.

A SELIC é a taxa básica de juros da economia definida pelo Banco Central do Brasil (BC). É um instrumento de política monetária, definidor do custo básico do crédito e utilizada como uma das ferramentas de controle da inflação (BANCO CENTRAL DO BRASIL, 2022b). É importante registrar que todas as dívidas inscritas em DAU são atualizadas monetariamente pela SELIC, por expressa determinação do artigo 30, da Lei nº 10.522, de 19 de julho de 2002.

O PIB é o medidor básico do crescimento econômico de um país, é o produto de toda a riqueza produzida em solo nacional (INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA, 2022b). Um PIB positivo e crescente aponta uma economia pujante.

Para além de serem alguns dos principais indicadores macroeconômicos, a escolha desses parâmetros se justifica, acima de tudo, porque suas projeções futuras são apuradas e divulgadas semanalmente pelo Banco Central do Brasil na publicação Relatório Focus (https://www.bcb.gov.br/publicacoes/focus). Logo, se o presente estudo busca indicar um caminho eficiente para a projeção da arrecadação, prudente se valer de indicadores cujas estimativas são calculadas pelos principais agentes da economia.

A arrecadação da DAU, entretanto, não decorre apenas de fatores externos, estranhos à atividade da Instituição responsável por sua cobrança. Há que se considerar, ainda, ações implementadas pelo Estado para ampliar a arrecadação.

De modo geral, a cobrança da dívida ativa é feita através de ações administrativas e do ajuizamento de execuções fiscais. A utilização de mecanismos de remissão e diferimento da dívida também são muitas vezes utilizados pela União para tentar ampliar a entrada de recursos. Tais medidas se dão, em sua maioria, pela instituição de parcelamentos extraordinários (PROCURADORIA-GERAL DA FAZENDA NACIONAL, 2021b). Mais recentemente, a Lei nº. 13.988/2020 regulamentou e autorizou à PGFN a implantação de medidas de transação fiscal, que podem ser simplificadas como negociações individualizadas com o contribuinte devedor para adoção de medidas e planos personalizados de pagamento (PROCURADORIA-GERAL DA FAZENDA NACIONAL, 2021c). Por essa razão, pertinente que essas ações também sejam consideradas na modelagem do estudo, dado seu potencial impacto na arrecadação.

2. METODOLOGIA

Segundo Sinan Ozdemir (2016, p. 202), o Aprendizado de Máquina ou Machine Learning se constitui numa ferramenta da Ciência de Dados preocupada com a capacidade de encontrar padrões de dados, mesmo que eles contenham erros e ruídos. Modelos de aprendizado de máquina são capazes de extrair conhecimento de amostras sem uma ajuda humana direta, o que lhes diferencia de algoritmos tradicionais. A maioria dos algoritmos de aprendizado de máquina focam em encontrar relacionamentos entre conjuntos de dados para identificar correlações que possam ser utilizadas para prever observações futuras.

Moreira, Carvalho e Horváth (2018) explicam que as tarefas de predição podem ser divididas em problemas de classificação e problemas de regressão. Na classificação, o objetivo primordial é que o algoritmo seja capaz de atribuir um rótulo (ex.: identificar, numa foto, se o que está representado é uma paisagem, um objeto, um animal etc.). Na regressão, busca-se a criação de uma função matemática que possa explicar uma saída a partir de um conjunto de atributos (ex.: estimar o tempo de viagem, dado um determinado caminho e as condições climáticas registradas).

Algoritmos de regressão têm sido usados em diversas áreas nos últimos anos, em alguns casos para estimar o valor futuro do mercado de ações (SANTOS, 2020), a evasão de clientes de um determinado produto (OLIVEIRA, 2021) e até a inflação da economia (FREITAS, 2019).

Portanto, para o objetivo pretendido, o mais adequado é o uso de algoritmos de regressão. Esse estudo é inovador em relação ao objeto, mas não quanto à técnica. Não se encontram trabalhos similares envolvendo a Dívida Ativa, mas os modelos utilizados são de amplo conhecimento, tendo sido testados em diversos conjuntos de dados, envolvendo áreas distintas.

Os arquivos das bases de dados utilizadas e os códigos de programação para treinamento e teste dos modelos estão disponíveis para consulta em repositório público no Github no endereço https://github.

com/search?q=estimativa_dau. Nos tópicos seguintes são detalhados os algoritmos escolhidos e a forma de obtenção e respectivas fontes dos dados.

2.1. Modelos de Machine Learning Utilizados

Existem diversos métodos de aprendizado de máquina para análise de dados automatizada que podem ser utilizados com fins preditivos.

Kelleher e Tierney (2018, p. 70) explicam que para algumas áreas há razões teóricas, baseadas no tipo particular de relacionamento entre os atributos e o valor que se pretende prever, que sugerem a aplicação de algoritmos específicos, todavia, na ausência desse tipo de teoria para o domínio em discussão, constitui uma boa prática assumir que a forma de relacionamento mais simples – o relacionamento linear – é suficiente, só então sendo pertinente progredir para modelos mais complexos.

Seguindo essa premissa, optou-se por avaliar o uso dos algoritmos de predição: *Regressão Linear (Linear Regression)*, *Árvore de Decisão (Decision Tree)*, *Floresta Randômica* ou *Floresta Aleatória (Random Forest)* e *Árvore de Decisão de Aumento de Gradiente (Gradient Boosting Decision Tree)*.

A *Regressão Linear* procura estimar um valor, dado um conjunto de atributos fixos, a partir da identificação de uma função de regressão (KELLEHER E TIERNEY, 2018, p. 70), ou seja, uma equação linear do tipo:

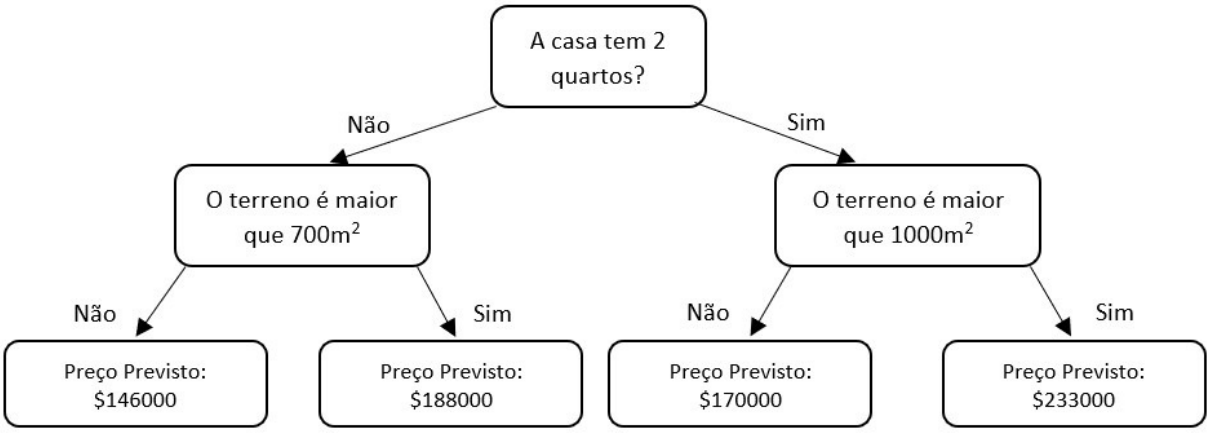
$$Y = w_0 + w_1X_1 + w_2X_2 + \ldots + w_aX_a$$

Basicamente, trata-se de uma equação que considera uma ou mais variáveis (X_1, X_2, ... X_a) e seus respectivos coeficientes paramétricos (w_0, w_1, w_2, ..., w_a) para encontrar o valor alvo (Y). O trabalho do algoritmo de aprendizado de máquina é, a partir do conjunto de dados, encontrar o coeficiente que influencia o peso de cada variável atributo da equação.

No modelo de *Árvore de Decisão*, para encontrar a fórmula preditiva são aplicadas uma série de classificações dos dados, camada a camada, de forma a melhor identificar a equação que se adequa aos atributos disponíveis (KELLEHER E TIERNEY, 2018, p. 82).

Utilizando o exemplo de Dan Becker (2017a), o modelo de árvore de decisão para previsão do valor de uma casa pode ser ilustrado da seguinte maneira:

FIGURA 1 - FUNCIONAMENTO DA ÁRVORE DE DECISÃO



Fonte: BECKER (2017a)

A última instância da árvore é denominada "folha" (BECKER, 2017a) e nela encontra-se o modelo preditivo que mais se adequa ao conjunto de dados, considerando as avaliações prévias feitas pelo algoritmo.

A *Floresta Randômica* faz uso de muitas árvores de decisão, cada uma treinada com um subconjunto dos dados disponibilizados para o algoritmo, a estimativa é feita calculando a média das previsões dessas várias árvores (KELLEHER E TIERNEY, 2018, p. 85).

Finalmente, a *Árvore de Decisão de Aumento de Gradiente* também faz uso de muitas árvores de decisão para conseguir um resultado satisfatório, mas de forma diferente do modelo *Floresta Randômica*. No GBDT (acrônimo para *Gradient Boosting Decision Tree*) as árvores de decisão são calculadas em sequência, cada uma levando em consideração o coeficiente de erro da árvore anterior, assim a precisão do modelo aumenta a cada nova árvore criada (ZHANG et al., 2021). Ao contrário da Floresta Randômica, em que as diversas árvores são calculadas de maneira independente e ao final a precisão é determinada pela ponderação da acurácia de cada uma das árvores geradas.

As breves explicações acima destinam-se a elucidar minimamente o funcionamento dos modelos escolhidos, porém não constitui escopo desse estudo analisar os teoremas matemáticos e a álgebra linear requerida para calcular e elaborar as previsões. Atualmente, com a popularização e simplificação dos modelos computacionais, é possível utilizar ferramentas gratuitas e de código aberto para executar essas análises.

Em razão do seu alto nível, simplicidade e ênfase na legibilidade (FURTADO, 2022), optou-se por gerar os modelos com a linguagem de programação *Python*, que tem na biblioteca de código aberto *Scikit-learn* todo o ferramental necessário para as projeções pretendidas (PEDREGOSA, 2011). O estudo se vale ainda da biblioteca *Pandas*, também gratuita e de código aberto, criada para a linguagem *Python*, focada na manipulação e análise estatística de dados (MCKINNEY, 2010).

Além disso, as simulações foram executadas no ambiente *Google Colaboratory*, serviço gratuito disponibilizando on-line para escrever e executar códigos de programação em ambiente de nuvem, que dispensa o uso de computadores com alta capacidade, já que todo o processamento é feito pelos servidores do Google (2022).

2.2. Levantamento, Tratamento e Exploração dos Dados

Esclarecidos os algoritmos, passemos a detalhar os dados e suas fontes. A análise proposta leva em conta os indicadores macroeconômicos projetados semanalmente no Relatório Focus divulgado pelo Banco Central do Brasil: IPCA, IGPM, SELIC, taxa de Câmbio e PIB. Juntamente com esses dados foram considerados os parcelamentos excepcionais e as transações de créditos tributários efetivados pela PGFN.

Os indicadores macroeconômicos foram obtidos no site do BC, na ferramenta *Sistema Gerenciador de Séries Temporais* – SGS (https://www3.bcb.gov.br/sgspub/). Especificamente, foram capturadas as planilhas de códigos: "4380 - PIB mensal - Valores correntes (R$ milhões) – Função: Variação Percentual"; "13522 - Índice nacional de preços ao consumidor - amplo (IPCA) - em 12 meses – Função: Linear"; "4189 - Taxa de juros - Selic acumulada no mês anualizada base 252 - % a.a. – Função: Linear"; "3698 - Taxa de câmbio - Livre - Dólar americano (venda) - Média de período – mensal - Função: Linear"; e "189 - Índice geral de preços do mercado (IGP-M) - Var. % mensal – Função: Linear".

Ao contrário dos demais índices que são apresentados em percentual, o dólar foi considerado em valor monetário. Embora fosse possível apurar a variação percentual dos respectivos períodos, apropriada sua consideração em moeda corrente porque esta é a forma como o Banco Central estima seu valor, ou seja, as previsões semanais são em reais (R$) e não percentuais.

Os valores arrecadados mensalmente pela dívida ativa, entre janeiro de 2015 e dezembro de 2021, e as quantidades de adesões aos parcelamentos e transações foram adquiridos diretamente da PGFN, através de solicitação via Lei de Acesso à Informação, pedido nº. 03005.291447/2022-40, pesquisável em http://www.consultaesic.cgu.gov.br/busca/.

Um fator peculiar em relação a esses diferimentos e benefícios concedidos aos contribuintes é que, em geral, eles são casuísticos, não existindo um padrão absoluto a ser seguido pelo legislador, que no momento de sua aprovação pode alterar a proposta apresentada pelo Poder Executivo, que por sua vez também não segue, de forma definitiva, padrões específicos. A listagem inicial obtida registra, no intervalo avaliado, um total de 43 tipos diferentes de modalidades de parcelamentos e transações, o que comprova essa diversidade:

TABELA 1 - PARCELAMENTOS ESPECIAIS E TRANSAÇÕES REGISTRADOS NO PERÍODO DE 2015 A 2021

1	PROIES	23	Transação excepcional - demais débitos
2	Parcelamento da recuperação judicial	24	Transação excepcional - débitos previdenciários
3	Parcelamento PACAL	25	Transação excepcional - crédito rural e fundiário
4	PROFUT	26	Transação excepcional - simples nacional
5	Parcelamento da lei nº. 12.810/2013	27	Transação na dívida ativa tributária de pequeno valor - demais débitos
6	Liquidação crédito rural e fundiário	28	Transação na dívida ativa tributária de pequeno valor - débitos previdenciários
7	Parcelamento de arrematação	29	Transação na dívida ativa tributária de pequeno valor - simples nacional
8	Parcelamento especial - simples nacional	30	Repactuação - transação excepcional - demais débitos
9	Programa de regularização tributária - PRT	31	Repactuação - transação excepcional - débitos previdenciários
10	Parcelamento especial débitos previdenciários dos entes federativos	32	Repactuação - transação excepcional - crédito rural e fundiário
11	Programa especial de regularização tributária - PERT - demais débitos	33	Repactuação - transação excepcional - simples nacional
12	Programa especial de regularização tributária - PERT - débitos previdenciários	34	Transação no contencioso tributário
13	Programa de recuperação tributária rural	35	Transação do setor de eventos - PERSE - demais débitos
14	Parcelamento especial de regularização tributária - lei 13.496/2017 - demais débitos	36	Transação do setor de eventos - PERSE - débitos previdenciários
15	Parcelamento da lei n. 13586, de 28 de dezembro de 2017	37	Transação do setor de eventos - PERSE - simples nacional
16	Programa especial de regularização tributária - PERT - débitos não quitados no parcelamento 12.996	38	Repactuação - transação do setor de eventos - PERSE - demais débitos
17	Programa especial de regularização tributária Simples Nacional - PERT SN	39	Repactuação - transação do setor de eventos - PERSE - débitos previdenciários
18	Parcelamentos diversos	40	Repactuação - transação do setor de eventos - PERSE - simples nacional
19	Parcelamento de honorários advocatícios	41	Transação para regularização fiscal do simples nacional
20	Transação - demais débitos	42	Transação - dívida ativa tributária de pequeno valor - simples nacional - edital 1/2022
21	Transação - débitos previdenciários	43	Parcelamento excepcional de débitos previdenciários para municípios - EC 113/2021
22	Transação individual		

Fonte: Elaborado pelo autor

Dado o objetivo proposto, de viabilizar, para o futuro, uma metodologia de estimativa para a arrecadação, considerar cada parcelamento e transação em si não é a melhor alternativa, uma vez que, pelo histórico observado, muitas dessas medidas não se repetirão *ipsis litteris*, ao passo que outras poderão surgir em seu lugar, com suas próprias peculiaridades.

Isso em mente, tornou-se essencial sistematizar um modelo de categorização, de forma a permitir uma fácil classificação dos diferimentos tributários. Avaliando a lista de medidas levadas a efeito pelo poder público, é possível constatar algumas características. Nota-se que parcelamentos e transações podem atingir um setor específico da economia ou abranger um número genérico de contribuintes. Mas mesmo os genéricos podem ter sua aplicação direcionada, incidindo apenas sobre as empresas classificadas como integrantes do Simples Nacional ou abarcarem apenas débitos definidos como de pequeno valor.

Assim, propõem-se uma lista fechada de classificação que abarca todos os parcelamentos e transações que já foram efetivados, ao mesmo tempo, é capaz de enquadrar futuras medidas. Concebeu-se um sistema de representação em três níveis, considerando

o tipo de medida (parcelamento [P] ou transação [T]), sua abrangência (setorial [S] ou geral [G]) e sua setorização.

No nível da setorização, em sendo uma medida de abrangência geral, subclassifica-se em ampla [A] – sem qualquer restrição para adesão; simples nacional [SN] – destinada apenas às empresas enquadradas nessa modalidade de tributação; ou débitos de pequeno valor [PN] – abarcando somente débitos considerados pequenos pelo Fisco.

Já na hipótese das medidas de natureza setorial, optou-se por utilizar o código de atividades econômicas do CNAE – Classificação Nacional de Atividades Econômicas, do Instituto Brasileiro de Geografia e Estatística, por se tratar de modelo bastante sedimentado e utilizado no Cadastro Nacional de Pessoas Jurídicas da Receita Federal do Brasil (MINISTÉRIO DA ECONOMIA, 2014).

O próprio código CNAE é dividido em níveis, sendo a "seção" — designada por letras de *A* até *U* — destinada a representar o setor da atividade econômica desenvolvido pela empresa (INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA, 2022c). Utilizando especificamente essa informação, obtém-se então a abrangência necessária à classificação para permitir, no futuro, enquadrar novos tipos de diferimento tributário que possam ser criados.

Dessa forma, todos os parcelamentos e transações foram ordenados com base em um código de três símbolos:

TABELA 2 - ESQUEMA DE CATEGORIZAÇÃO DOS PARCELAMENTOS E TRANSAÇÕES

1º Símbolo	Tipo:	[P] parcelamento ou [T] transação
2º Símbolo	Abrangência:	[S] setorial ou [G] geral
3º Símbolo	Setorização:	[SN] simples nacional, [PV] pequeno valor, [A] amplo ou [seção] código CNAE

Fonte: Elaborado pelo autor

O PROIES, por exemplo, sendo um parcelamento destinado a atender as empresas do setor de ensino, foi enquadrado no código P-S-P. A primeira letra referente ao tipo de medida (P – parcelamento), a segunda letra representando sua abrangência (S – setorial) e a terceira letra a seção CNAE a que pertence (P – educação). As transações viabilizadas para as empresas enquadradas no Simples Nacional, ao seu turno, receberam o código T-G-SN, ou seja, [T] de transação, [G] por ter abrangência geral e [SN] porque se destina apenas aos débitos do Simples Nacional.

Por essa sistemática é possível a criação de um total de 47 códigos, com os registros de todos os parcelamentos e transações detalhadas até o nível do setor da economia beneficiado. Contudo, grande parte dessas classes servem para o futuro, mas não possuem registros passados para alimentar os modelos. Assim, apenas as 12 classes que registraram dados, entre 2015 e 2021, foram usadas como atributo para os algoritmos de regressão:

TABELA 3 - CÓDIGO DE ENQUADRAMENTO DOS PARCELAMENTOS E TRANSAÇÕES

CÓD.	CATEGORIZAÇÃO
P-S-A	Parcelamento Setorial Agronegócio
P-S-B	Parcelamento Setorial Indústrias Extrativas
P-S-O	Parcelamento Setorial Administração Pública
P-S-P	Parcelamento Setorial Educação
P-S-R	Parcelamento Setorial Artes, Cultura, Esporte e Recreação
P-G-SN	Parcelamento Geral Simples Nacional
P-G-A	Parcelamento Geral Amplo
T-G-A	Transação Geral Ampla

CÓD.	CATEGORIZAÇÃO
T-G-SN	Transação Geral Simples Nacional
T-G-PV	Transação Geral Pequeno Valor
T-S-A	Transação Setorial Agronegócio
T-S-R	Transação Setorial Artes, Cultura, Esporte e Recreação

Fonte: Elaborado pelo autor

Foram agrupados em cada código os seguintes parcelamentos e transações:

TABELA 4 - AGRUPAMENTO DOS PARCELAMENTOS E TRANSAÇÕES POR CÓDIGO

CÓD.	PARCELAMENTOS OU TRANSAÇÕES AGRUPADOS
P-S-A	Parcelamento PACAL;
	Liquidação Crédito Rural e Fundiário;
	Programa de recuperação tributária rural;
P-S-B	Parcelamento da lei nº. 13.586/2017;
P-S-O	Parcelamento de Recuperação Judicial;
	Parcelamento da LEI nº. 12.810/2013;
	Parcelamento de Arrematação;
	Parcelamento Especial Débitos Previdenciários dos Entes Federados;
	Parcelamento de honorários advocatícios;
	Parcelamento excepcional de débitos previdenciários para municípios - EC 113/2021;
P-S-P	PROIES;
P-S-R	PROFUT;
P-G-SN	Parcelamento Especial – Simples Nacional;
	Programa especial de regularização tributária Simples Nacional - PERT SN;
P-G-A	Programa de Regularização Tributária – PRT;
	Programa especial de regularização tributária - PERT - demais débitos;
	Programa especial de regularização tributária - PERT - débitos previdenciários;
	Parcelamento especial de regularização tributária - lei 13.496/2017 - demais débitos;
	Programa especial de regularização tributária - PERT - déb. não quitados no parc. 12.996;
	Parcelamentos diversos;
T-G-A	Transação - demais débitos;
	Transação - débitos previdenciários;
	Transação individual;
	Transação excepcional - demais débitos;
	Transação excepcional - débitos previdenciários;
	Repactuação - transação excepcional - demais débitos;
	Repactuação - transação excepcional - débitos previdenciários;
	Transação no contencioso tributário;
T-G-SN	Transação excepcional - simples nacional;
	Repactuação - transação excepcional - simples nacional;
	Transação para regularização fiscal do simples nacional;

CÓD.	PARCELAMENTOS OU TRANSAÇÕES AGRUPADOS
T-G-PV	Transação na dívida ativa tributária de pequeno valor - demais débitos;
	Transação na dívida ativa tributária de pequeno valor - débitos previdenciários;
	Transação na dívida ativa tributária de pequeno valor - simples nacional;
	Transação - dívida ativa tributária de pequeno valor - simples nacional - edital 1/2022;
T-S-A	Transação excepcional - crédito rural e fundiário;
	Repactuação - transação excepcional - crédito rural e fundiário;
T-S-R	Transação do setor de eventos - PERSE - demais débitos;
	Transação do setor de eventos - PERSE - débitos previdenciários;
	Transação do setor de eventos - PERSE - simples nacional;
	Repactuação - transação do setor de eventos - PERSE - demais débitos;
	Repactuação - transação do setor de eventos - PERSE - débitos previdenciários;
	Repactuação - transação do setor de eventos - PERSE - simples nacional;

Fonte: Elaborado pelo autor

Esses atributos foram tratados como variáveis binárias, ou variáveis *dummy*, e nos meses em que havia registro de parcelamento ou transação disponível para adesão pelos contribuintes, de acordo com o código de classificação proposto, foi atribuído o valor 1, não estando disponível, recebeu o valor 0. Essa opção se deve ao fato de que a consideração da quantidade de contribuintes que aderiram a esses parcelamentos constitui uma explicação qualitativa e não quantitativa, pois não há uma relação perfeita entre a quantidade de aderentes e o valor total arrecadado pela DAU. O parcelamento de um único débito de uma grande empresa, como uma petrolífera ou uma instituição financeira, pode gerar o ingresso de um montante financeiro muito superior que o de milhares de contribuintes pessoas físicas. Levar em conta a quantidade de contribuintes aderentes poderia representar uma ponderação arbitrária, visto que estariam sendo tratadas igualmente variáveis em escalas quantitativas e qualitativas, o que é considerado um erro nas avaliações estatísticas (FÁVERO E BELFIORE, 2017, p. 541).

Ademais, sob a perspectiva da elaboração de projeções futuras, a PGFN não tem como precisar a quantidade de pessoas que irão aderir a um determinado parcelamento ou transação, pode apenas dizer se estará disponível num determinado mês.

Os modelos preditivos foram alimentados, então, com 17 atributos (IPCA, IGPM, SELIC, Taxa de Câmbio, PIB e 12 códigos de parcelamentos e transações) e tiveram como valor alvo o montante da arrecadação mensal da Dívida Ativa da União.

Considerando que os dados apresentavam métricas diferentes (percentual e valores absolutos), para que essa diferença não impactasse na previsão foi aplicada uma função de normalização, mais especificamente o método *min-max*, em que cada valor dos atributos é subtraído do valor mínimo do respectivo conjunto de dados e o resultado encontrado é dividido pela diferença entre o maior e o menor valor do conjunto (amplitude), isso traz todos os indicadores para valores entre 0 e 1 e nenhuma variável apresenta maior peso nos algoritmos somente em razão da discrepância de unidade de medida (MOREIRA *et al.*, 2018, p. 91).

Sumarizando os dados com métodos de estatística descritiva[3], o período analisado é composto por 84 (oitenta e quatro) meses de registros, durante o qual as variáveis macroeconômicas consideradas apresentam as seguintes características:

3. Informações geradas com a função ".describe" da biblioteca Pandas. https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.describe.html

TABELA 5 – ESTATÍSTICA DESCRITIVA DAS VARIÁVEIS MACROECONÔMICAS - 2015 A 2021

	IPCA	IGP-M	SELIC	CÂMBIO	PIB
Média	5,74%	0,82%	8,17%	R$ 4,02	0,59%
Desvio Padrão	2,85%	1,05%	4,29%	R$ 0,88	3,76%
Menor Valor	1,88%	-1,10%	1,90%	R$ 2,63	-10,80%
Maior Valor	10,74%	4,34%	14,15%	R$ 5,65	10,50%
1º Quartil	3,26%	0,20%	4,98%	R$ 3,26	-1,53%
2º Quartil	4,56%	0,65%	6,40%	R$ 3,80	0,35%
3º Quartil	8,55%	1,19%	12,90%	R$ 4,92	2,68%

Fonte: Elaborado pelo autor

Neste mesmo período, a arrecadação média mensal da dívida ativa foi de R$ 1.945.842.649,48, com um desvio padrão de R$ 801.734.014,24. A menor arrecadação em um único mês foi de R$ 1.079.880.695,58, a maior R$ 5.705.095.119,62 e o primeiro, segundo e terceiro quartis correspondentes a R$ 1.314.979.048,29, R$ 1.774.879.257,38 e R$ 2.255.027.522,03, respectivamente.

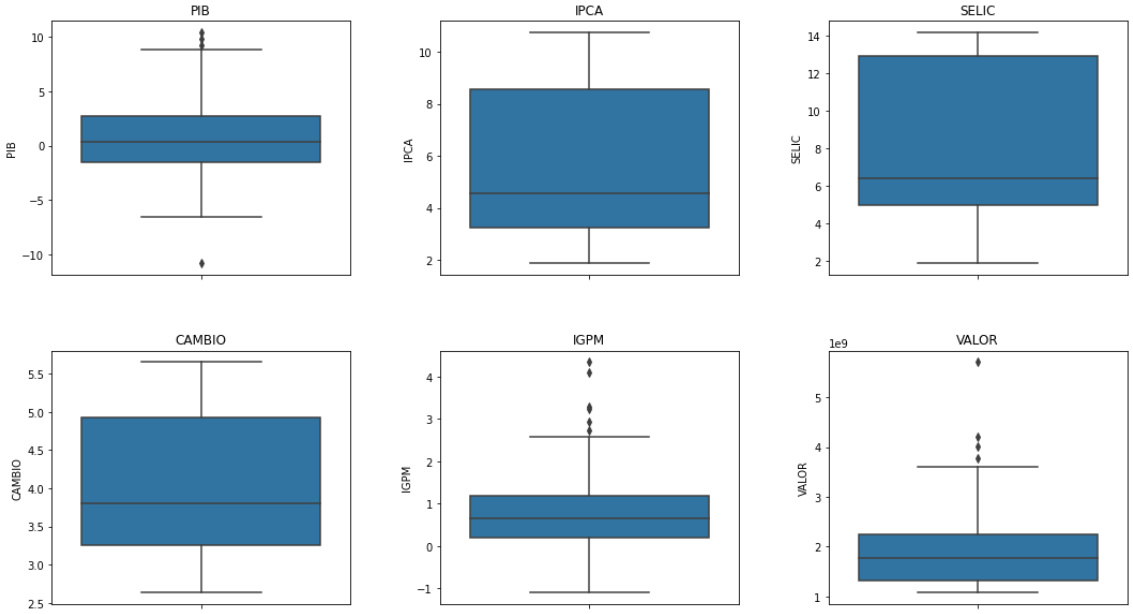
Fávero e Belfiori (2017, p. 53 a 55) alertam que:

> Um conjunto de dados pode conter algumas observações que apresentam um grande afastamento das restantes ou são inconsistentes. Estas observações são designadas por *outlíers*, ou ainda por valores atípicos, discrepantes, anormais ou extremos.
>
> Antes de decidir o que será feito com as observações outliers, devemos ter o conhecimento das causas que levaram a tal ocorrência. Em muitos casos, essas causas podem determinar o tratamento adequado dos respectivos *outliers*. As principais causas estão relacionadas a erros de medição, de execução e variabilidade inerente aos elementos da população.
>
> (...)
>
> Caso seja identificado apenas um *outlier* em determinada variável, o pesquisador poderá tratá-lo por meio de alguns procedimentos existentes, por exemplo, a eliminação completa desta observação. Por outro lado, se houver mais de um *outlier* para uma ou mais variáveis individualmente, a exclusão de todas as observações pode gerar uma redução significativa do tamanho da amostra. (...).
>
> (...), se a intenção for justamente analisar o comportamento dessas observações atípicas ou de criar subgrupos por meio de critérios de discrepância, talvez a eliminação dessas observações ou a substituição dos seus valores não seja a melhor solução.

Pela identificação da amplitude interquartil, que corresponde à diferença entre o terceiro e o primeiro quartil do conjunto de dados, foram considerados outliers os valores posicionados acima do terceiro quartil ou abaixo do primeiro quartil por uma margem superior a uma vez e meia a amplitude (FÁVERO E BELFIORE, 2017, p. 53).

Como mostram os gráficos do tipo *boxplot* abaixo, o conjunto de dados possui *outliers* na arrecadação (quatro registros superiores a R$ 3,665 bi); no IGPM (seis registros superiores a 2,68%); e no PIB (três registros superiores a 9% e um registro inferior a -7,85%):

FIGURA 2 - BOXPLOT ATRIBUTOS QUANTITATIVOS



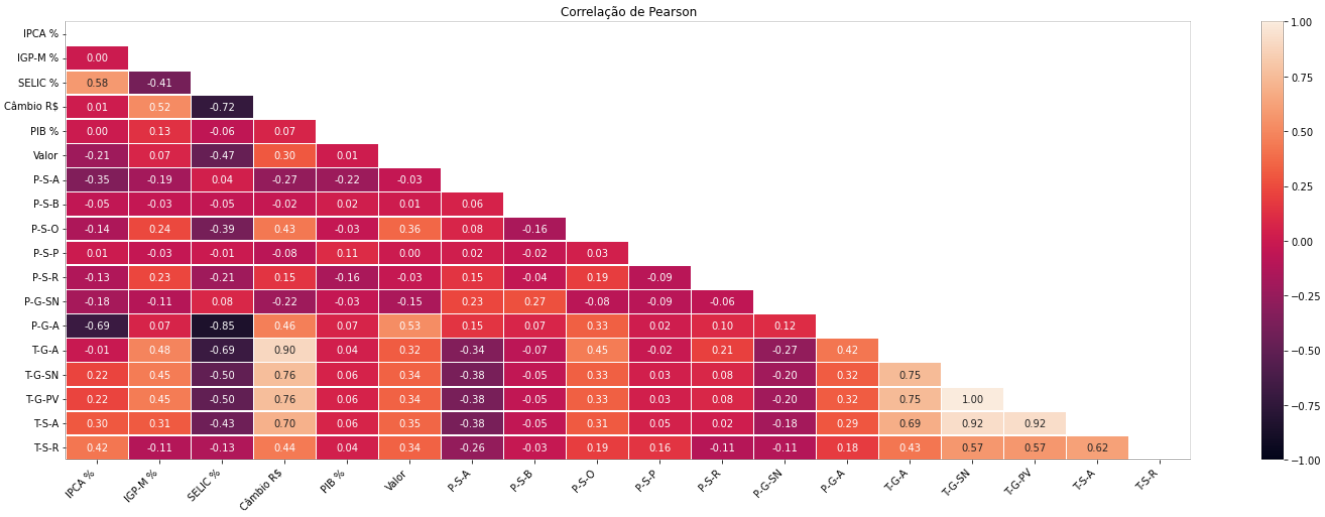
Fonte: Elaborado pelo autor

Esses valores, porém, não receberam qualquer tipo de tratamento, primeiro por se mostrarem registros válidos, constatações do mercado real, que apesar de incomuns não podem ser considerados equivocados; segundo, em razão da base de dados possuir uma quantidade limitada de informações, somente 84 meses, para a desconsideração desses valores seria necessária a exclusão de 14 meses da amostra, uma redução significativa e com provável impactando na precisão das estimativas.

Por derradeiro, calculou-se o *coeficiente de correlação de Pearson*[4], que verifica o tipo de relação linear das variáveis. Sua medida varia entre -1 e 1, a aproximação dos extremos está vinculada a um aumento da intensidade da correlação, o sinal indica se a relação é direta ou inversa e quanto mais próximo de zero, menor a correlação (FÁVERO E BELFIORE, 2017, p. 118).

4. Coeficientes apurados com a função ".corr" do Pandas. https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.corr.html

FIGURA 3 - COEFICIENTE DE PEARSON



Fonte: Elaborado pelo autor

Seguindo o critério de avaliação qualitativa do grau de correlação entre as variáveis apresentado por Callegari-Jacques (2007, p. 90), em que um coeficiente 0 representa uma correlação nula; entre 0 e 0,3, fraca; acima de 0,3 até 0,6, regular; acima de 0,6 até 0,9, forte; acima de 0,9 até 1, muito forte; e 1 significa uma correlação plena ou perfeita; é possível afirmar que das variáveis macroeconômicas estudadas, a SELIC é a que apresenta uma correlação mais acentuada com a arrecadação (igual a -0,47), de grau regular e natureza inversa, isto é, o aumento da arrecadação está associado à redução da taxa básica de juros da economia. Câmbio, IGP-M e IPCA têm correlações fracas com a arrecadação (0,30, 0,07 e -0,21, respectivamente). Já com o PIB parece não existir qualquer correlação (0,01).

A visualização da correlação dos parcelamentos e transações com a arrecadação, ao seu turno, leva a constatações importantes. Enquanto todas as modalidades de transação apresentaram correlações regulares diretas com os movimentos da arrecadação, os parcelamentos excepcionais possuem comportamentos conflitantes. Há uma relação de grau regular entre os valores dos ingressos e os parcelamentos gerais amplos (0,53), os parcelamentos setoriais, contudo, não aparentam possuir relações significativas com a arrecadação, com exceção do parcelamento destinado ao setor público (P-S-O), que demonstrou um grau regular de correlação. O parcelamento geral destinado aos débitos do simples nacional (P-G-SN) chegou a apresentar uma correlação negativa (-0,15), embora de grau fraco.

É fundamental lembrar, porém, que correlação não significa causalidade. O coeficiente de Pearson aponta o sentido e a intensidade com que as variáveis caminham na comparação mútua, mas não é possível afirmar, sem uma pesquisa mais aprofundada, o grau de determinação entre elas. Como alerta Paranhos (2014), acerca do perigo das correlações espúrias, dada a complexidade das interações sociais é difícil avaliar um fenômeno exclusivamente por meio de sua relação com uma única outra variável.

De todo modo, a partir dos dados analisados já é possível questionar o senso comum que atribui a todo e qualquer tipo de parcelamento excepcional uma ampliação da arrecadação federal, o que é comumente verbalizado quando das discussões desses tipos de diferimento no Congresso Nacional (AGÊNCIA SENADO, 2021). Neste sentido, as transações tributárias instituídas pela PGFN se mostraram mais consistentes, o que sugere que uma avaliação dos pontos convergentes e divergentes entre os dois institutos pode trazer insights para a formulação de políticas públicas mais profícuas nesta área.

3. RESULTADOS DOS MODELOS

Preliminarmente, salutar esclarecer que os modelos foram avaliados utilizando-se a técnica do *coeficiente de ajuste* R^2, que considera o quadrado da soma das distâncias entre os dados observados e a média desses dados (SQt), subtraído pelo quadrado da soma das distâncias entre os dados observados e os dados gerados pelo modelo de aprendizado de máquina (SQr). O número encontrado é dividido pelo quadrado da soma das distâncias entre os dados observados e a média desses dados (SQt), o que resulta num valor entre 0 e 1 (FÁVERO E BELFIORE, 2017, p. 522).[5]

O valor encontrado pode ser transformado em percentual bastando multiplicá-lo por 100. Esse percentual indica o quanto o modelo explica a variação dos dados, ou seja, um percentual de 75%, por exemplo, significa que o modelo preditivo explica a variação dos dados em 75% das ocorrências, ou que é 75% mais preciso que simplesmente calcular a média dos valores observados.

Além disso, apurou-se o Erro Absoluto Médio (*Mean Absolute Error* - MAE)[6], que é obtido pelo cálculo da média das diferenças absolutas entre os valores previstos pelo modelo e o valor correto (BECKER, 2017b). Portanto, no nosso caso, um MAE de 200.000.000 por exemplo significa que as previsões feitas pelo modelo estão, em média, até R$ 200 mi distante do valor correto.

O treinamento dos modelos foi feito utilizando 75% da amostra (carga de treino) e a validação dos resultados considerou os 25% restantes (carga de teste). A separação dos dados entre as duas cargas se deu por meio de função automática da biblioteca *Scikit-learn*[7], de modo randômico. A tabela a seguir condensa os melhores resultados obtidos:

TABELA 6 – RESULTADOS DOS TESTES DOS MODELOS

MODELO	R^2		ERRO MÉDIO ABSOLUTO
Regressão Linear	0,5228	52,28%	342.584.890,21
Árvore de Decisão	0,9323	93,23%	144.108.108,24
Floresta Randômica	0,8306	83,06%	119.490.899,62
Árvore de Decisão de Aumento de Gradiente	0,9476	94,76%	117.320.076,32

Fonte: Elaborado pelo autor (2022)

O modelo de Árvore de Decisão de Aumento de Gradiente (*Gradient Boosting Decision Tree*) se mostrou o mais bem sucedido em prever os valores arrecadados, sendo capaz de explicar a variação da arrecadação mensal em 94,76% das vezes e apresentou uma divergência média, para mais ou para menos, entre o valor de fato arrecadado e o previsto de aproximadamente R$ 117,3 mi.

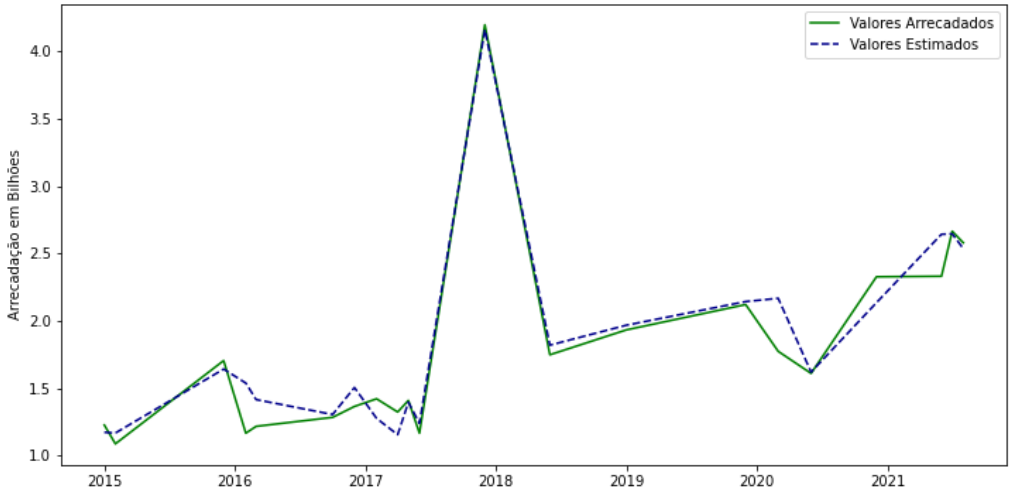
Considerando os 21 meses de dados utilizados na carga de teste, o total arrecadado pela DAU foi de R$ 37.604.327.956,75, o total previsto pelo modelo, para os mesmos registros, foi de R$ 38.634.416.006,69, uma diferença percentual de aproximadamente 2,57%. A comparação gráfica entre os valores previstos e o real se apresenta da seguinte maneira:

5. O cálculo dos coeficientes de ajuste R2 foi efetuado de maneira automatizada, com a função "r2_score" da biblioteca Scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html

6. Os erros médios foram obtidos de maneira automatizada, utilizando a função "mean_absolute_error" da biblioteca Scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_absolute_error.html

7. Função "train_test_split". https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

FIGURA 4 - GRÁFICO RESULTADO ÁRVORE DE DECISÃO DE AUMENTO DE GRADIENTE



Fonte: Elaborado pelo autor (2022)

A título de comparação, valendo-se da suavização exponencial a PGFN, em suas últimas quatro estimativas, obteve uma diferença média entre o previsto e o real de 9,46%. Em sua melhor previsão, a divergência foi de 4,2% (PROCURADORIA-GERAL DA FAZENDA NACIONAL, 2021a).

4. CONCLUSÃO

Tendo como parâmetro de comparação os índices de acerto das estimativas de arrecadação realizadas pela PGFN utilizando o método de suavização exponencial, é possível afirmar que o uso dos algoritmos de *Machine Learning* Árvore de Decisão (*Decision Tree*) e Árvore de Decisão de Aumento de Gradiente (*Gradient Boosting Decision Tree*) mostraram-se bem-sucedidos, obtendo desempenho superior em prever os movimentos de ingressos da Dívida Ativa, quando alimentados com dados de indicadores macroeconômicos (PIB, IPCA, IGP-M, taxa de Câmbio e taxa SELIC) e das estratégias de arrecadação adotadas pela União (parcelamentos excepcionais e transações tributárias).

O modelo GBDT, além de apresentar o melhor coeficiente de previsão (coeficiente de ajuste R2 = 94,76%), também teve a menor média de erro entre o valor previsto e o valor de fato apurado (erro médio absoluto = 117.320.076,32). Para os 21 meses de dados que compuseram a carga de teste dos modelos, a diferença entre a soma dos valores previstos pela Árvore de Decisão de Aumento de Gradiente e a soma dos valores de fato arrecadados foi de 2,57%.

Todavia mais importante que os índices de acerto dos modelos, sua utilização revela-se mais vantajosa por levar em consideração, para a estimativa de ingresso de valores, variáveis que atualmente são desprezadas. Isso significa que os modelos propostos estão ancorados em parâmetros mais amplos e robustos, enquanto os cálculos atuais refletem somente os registros passados da própria arrecadação.

Ao elaborar as projeções de receita para a formulação da Lei Orçamentária Anual, a PGFN, com a metodologia proposta, poderá levar em conta os dados que os agentes econômicos preveem para a economia nacional, o que é mais consistente com as expectativas financeiras nacionais.

Outro ganho que sobressai está na possibilidade da Administração utilizar os algoritmos de Machine Learning para avaliar, a priori, os impactos arrecadatórios da instituição de parcelamentos e transações tributárias, o que viabiliza uma decisão pública mais bem amparada e pautada em fatores empiricamente simulados e testados.

Há que se observar, ainda, que uma análise descritiva dos dados demonstra eficácia maior das transações tributárias em detrimento dos parcelamentos excepcionais, ao menos do modo como veem sendo instituídos nos últimos seis anos, o que merece maiores estudos, notadamente visando identificar as razões desses resultados diversos com vista a aprimorar a construção de políticas públicas eficientes.

REFERÊNCIAS

AGÊNCIA SENADO. (2021). *Projeto que reabre prazo para o Programa Especial de Regularização Tributária segue para a Câmara*. Brasília, BR. https://www12.senado.leg.br/noticias/materias/2021/08/05/projeto-que-reabre-prazo-para-o-programa-especial-de-regularizacao-tributaria-segue-para-a-camara

BANCO CENTRAL DO BRASIL. (2022a). *Estatísticas do setor externo*. Nota para a imprensa - 29/04/2022. https://www.bcb.gov.br/estatisticas/estatisticassetorexterno

_____. (2022b). Taxa Selic. https://www.bcb.gov.br/controleinflacao/taxaselic

BECKER, Dan. (2017a). *How Models Work: The first step if you're new to machine learning*. Kaggle. https://www.kaggle.com/code/dansbecker/how-models-work

_____. (2017b). *Model Validation: Measure the performance of your model, so you can test and compare alternatives*. Kaggle. https://www.kaggle.com/code/dansbecker/model-validation

CALLEGARI-JACQUES, Sidia M. (2017). *Bioestatística: princípios e aplicações*. Porto Alegre, BR: Artmed.

FÁVERO, Luiz Paulo; BELFIORE, Patrícia. (2017). *Manual de análise de dados*. 1. ed. Rio de Janeiro, BR: Elsevier.

FUNDAÇÃO GETÚLIO VARGAS. (2022). *IGP-M: Resultados 2022*. FGV. https://portal.fgv.br/noticias/igpm-resultados-2022

FURTADO, Paulo Augusto. (2022). Por que escolher Python?. In: *Jornada Python: uma jornada imersiva na aplicabilidade de uma das mais poderosas linguagens de programação do mundo*. Rio de Janeiro, BR: Brasport.

FREITAS, Gabriel Belmino. (2019). *O uso de machine learning na modelagem da previsão de inflação: revisão bibliográfica*. 42 f. Trabalho de Conclusão de Curso (Bacharelado em Ciências Econômicas) — Universidade de Brasília, Brasília, 2019. https://bdm.unb.br/handle/10483/25328

GOOGLE. (2022). *Conheça o Colab*. https://colab.research.google.com/?utm_source=scs-index#scrollTo=OwuxHmxllTwN

GRUS, Joel. (2016). *Data Science do Zero*. Traduzido por Welington Nascimento. Rio de Janeiro, BR: Alta Books.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. (2022a). *IPCA - Índice Nacional de Preços ao Consumidor Amplo*. Portal do Governo Brasileiro. https://www.ibge.gov.br/estatisticas/economicas/precos-e-custos/9256-indice-nacional-de-precos-ao-consumidor-amplo.html?=&t=o-que-e

____. (2022b). *Produto Interno Bruto - PIB*. Portal do Governo Brasileiro. https://www.ibge.gov.br/explica/pib.php

____. (2022c). *CONCLA - Comissão Nacional de Classificação*. https://cnae.ibge.gov.br/?view=estrutura&tipo=cnae&versao_classe=7.0.0&versao_subclasse=9.1.0

KELLEHER, John D.; TIERNEY, Brendan. (2018). *Data Science. The MIT Press essential knowledge series*. Cambridge, US: The MIT Press.

KRUGMAN, Paul R; OBSTFELD, Maurice. (2005). *Economia internacional: teoria e política*. Tradutor técnico Eliezer Martins Diniz. São Paulo, BR: Pearson Addison Wesley.

MCKINNEY, Wes. (2010). Data structures for statistical computing in python, *Proceedings of the 9th Python in Science Conference*, Volume 445. https://conference.scipy.org/proceedings/scipy2010/pdfs/mckinney.pdf

MINISTÉRIO DA ECONOMIA. (novembro, 2014). *Cadastro Nacional de Atividades Econômicas – CNAE*. https://www.gov.br/receitafederal/pt-br/assuntos/orientacao-tributaria/cadastros/cnpj/classificacao-nacional-de-atividades-economicas-2013-cnae/apresentacao

MOREIRA, J. M.; CARVALHO, A.; HORVÁTH, T. (2018). *A general introduction to data analytics*. Hoboken, US: Wiley.

OLIVEIRA, B. (2021). *Algoritmos de aprendizado de máquina na predição e avaliação de evasão de clientes em ambiente de produção*. 87 f. Dissertação (Mestrado em Ciência da Computação) - Universidade Federal de Goiás. http://repositorio.bc.ufg.br/tede/handle/tede/11522

OZDEMIR, Sinan. (2016). *Principles of data science*. Birmingham, UK: Packt Publishing Ltd.

PARANHOS, R. et al. (2014). Desvendando os Mistérios do Coeficiente de Correlação de Pearson: o Retorno. *Leviathan (São Paulo)*, (8), 66-95. DOI: https://doi.org/10.11606/issn.2237-4485.lev.2014.132346

PEDREGOSA *et al*. (2011). *Scikit-learn: Machine Learning in Python*, Journal of Machine Learning Research 12, pp. 2825-2830. https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

PROCURADORIA-GERAL DA FAZENDA NACIONAL. (2021a). Nota SEI nº 29/2021/PGDAU-CDA-COAGED/PGDAU-CDA/PGDAU/PGFN-ME. *Presta informações sobre a dívida ativa da União para compor o anexo de riscos fiscais*. Ministério da Economia. http://www.consultaesic.cgu.gov.br/busca/dados/Lists/Pedido/Attachments/1637685/RESPOSTA_RECURSO_1_161568_SEI_ME___14626293___Nota.pdf

_____. (2021b). Nota SEI nº 7/2022/PGDAU-CDA-COAGED/PGDAU-CDA/PGDAU/PGFN-ME. *Boletim de Acompanhamento Gerencial - Edição Anual – 2021*. Ministério da Economia. https://www.gov.br/pgfn/pt-br/assuntos/divida-ativa-da-uniao/estudos-sobre-a-dau/boletim-de-acompanhamento-gerencial-da-divida-ativa-da-uniao-e-do-fgts-edicao-anual-2021.pdf

_____. (2021c). Nota Conjunta SEI nº 2/2021/PGDAU-CGR. *Analisa os resultados alcançados pelas modalidades de transação da dívida ativa da União e da transação do contencioso de pequeno valor*. Ministério da Economia. https://www.gov.br/pgfn/pt-br/assuntos/divida-ativa-da-uniao/estudos-sobre-a-dau/sei_me-17016922-nota-conjunta.pdf

SANTOS, Gustavo Carvalho. (2020). *Algoritmos de Machine Learning para previsão da B3*. 90 f. Dissertação (Mestrado em Engenharia Elétrica) - Universidade Federal de Uberlândia, Uberlândia. DOI https://doi.org/10.14393/ufu.di.2020.640

ZHANG, Z. *et. al*. (julho, 2021). GBDT-MO: *Gradient-Boosted Decision Trees for Multiple Outputs* in IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 7, pp. 3156-3167. http://proceedings.mlr.press/v70/si17a.html

Rubens Quaresma Santos
E-mail: rubens.quaresma@pgfn.gov.br
Instituição de vinculação: Procuradoria-Geral da Fazenda Nacional - PGFN
ID ORCID: https://orcid.org/0000-0002-3170-4348

Pós-graduado (lato sensu) em Administração Pública pela Fundação Getúlio Vargas (2016). Graduado em Direito pela Universidade Estadual de Feira de Santana (2006). Graduando em Engenharia de Software pelo Instituto de Educação Superior de Brasília. Procurador na Procuradoria-Geral da Fazenda Nacional (PGFN), tendo exercido as funções de Subprocurador-Regional da Fazenda Nacional na 1ª Região (2017-2019) e de Procurador-Regional da Fazenda Nacional na 1ª Região (2019-2021), atualmente integra a Divisão de Consultoria e Assessoramento Jurídico da Procuradoria-Regional da Fazenda Nacional na 1ª Região.