



Lisbon School
of Economics
& Management
Universidade de Lisboa

TRABALHO FINAL DO MESTRADO

PROJETO

INTERPRETABILIDADE DE MODELOS MACHINE LEARNING PARA A PREVISÃO MACROECONÓMICA

MARIA A. CARMO REBELO

ORIENTAÇÃO:

PROFESSOR DOUTOR JOÃO AFONSO BASTOS

**DOCUMENTO ESPECIALMENTE ELABORADO PARA OBTENÇÃO DO GRAU DE MESTRE
EM MÉTODOS QUANTITATIVOS PARA A DECISÃO ECONÓMICA E EMPRESARIAL**

ISEG, OUTUBRO 2022

RESUMO

A interpretabilidade de um modelo pode ser definida como o grau de compreensão sobre o funcionamento interno de um modelo, permitindo perceber as causas sobre o qual recai o resultado. Na modelação estatística, pode verificar-se um *trade-off* entre a interpretabilidade e a precisão da previsão: modelos de interpretabilidade natural revelam uma menor precisão relativamente a modelos mais complexos. A previsão do Produto Interno Bruto é um aspeto fundamental para a aplicação de políticas económicas, com particular atenção para políticas monetárias e políticas orçamentais. Nesse sentido e tendo em conta que os modelos *de machine learning* preveem melhor a volatilidade económica, tem se vindo a assistir a uma revolução do método de previsão macroeconómico utilizado pelas principais organizações políticas nacionais e internacionais.

O objetivo deste trabalho prende-se em investigar se um *black-box*, *Extreme Gradient Boosting* (XGBoost) pode superar os métodos de interpretabilidade natural selecionados, Regressão linear e Árvore de decisão na previsão do Produto Interno Bruto (PIB) *per capita* para dados de painel e identificar medidas de importância das variáveis para melhorar a transparência dos modelos de *machine learning*. Se as organizações políticas e financeiras forem capazes de prever e interpretar corretamente os fatores, podem implementar políticas mais eficazes.

Para a análise de interpretabilidade de XGBoost, utilizam-se os seguintes métodos, *Shapley Additive ExPlanations* (SHAP), *Permutation Feature Importance* e *Partial Dependence Plot*. Através destes métodos pretendemos mostrar que os resultados obtidos por XGBoost podem ser interpretados sem um grande esforço computacional e garantir maior vantagem competitiva. Contudo, como será aprofundado durante este trabalho, verifica-se que o modelo XGBoost será o modelo com melhor precisão de previsão para os dados de painel.

PALAVRAS-CHAVE: Produto Interno Bruto *per capita*; Dados de painel; *Time-fixed effects*; XGBoost; Interpretabilidade de modelos.

ABSTRACT

The interpretability of a model can be defined as the degree of understanding of the inner workings of a model, allowing one to understand the causes upon which the result is based. In statistical modelling, there can be a trade-off between interpretability and forecast accuracy: models with natural interpretability show lower accuracy than more complex models. Gross Domestic Product forecasting is a fundamental aspect for the implementation of economic policies, with particular attention to monetary policies and fiscal policies. In this sense and considering that machine learning models better predict economic volatility, there has been a revolution in the macroeconomic forecasting method used by the main national and international political organizations.

The objective of this work is to investigate whether a black-box, Extreme Gradient Boosting (XGBoost) can outperform selected natural interpretability, Linear regression and Decision tree methods in forecasting Gross Domestic Product (GDP) per capita for panel data and to identify measures of variable importance to improve the transparency of machine learning models. If political and financial organizations can correctly predict and interpret factors, they can implement more effective policies.

For the XGBoost interpretability analysis, the following methods are used, Shapley Additive ExPlanations (SHAP), Permutation Feature Importance and Partial Dependence Plot. Through these methods we intend to show that the results obtained by XGBoost can be interpreted without a large computational effort and ensure greater competitive advantage. However, as will be deepened during this work, it is verified that the fixed effects model will be the model with the best prediction accuracy for panel data.

KEYWORDS: Gross Domestic Product *per capita*; Panel data; Time-fixed effects; XGBoost; Model interpretability.

ÍNDICE

Resumo	i
Abstract.....	ii
Índice de figuras.....	iv
Índice de tabelas.....	vi
Agradecimentos	vii
1. Introdução	1
2. Revisão de literatura	2
2.1. Modelação da previsão macroeconómica.....	2
2.2. Indicadores económicos	4
2.3. Abordagens à interpretabilidade.....	6
3. Metodologia	8
3.1. Técnicas de modelação.....	8
3.1.1. Modelos de interpretabilidade natural.....	8
3.1.2. Black-box model: eXtreme Gradient Boosting (XGBoost).....	14
3.2. Abordagens de interpretabilidade seleccionadas.....	18
3.2.1. Classical Feature Importance	18
3.2.2. Permutation Feature Importance.....	18
3.2.3. Partial Dependence Plots (PDP)	18
3.2.4. SHAP (SHapley Additive ExPlanations) Values.....	19
3.3. Amostragem de treino e teste	20
3.4. Avaliação de precisão preditiva.....	21
4. Dados	22
4.1. Index of economic freedom database	22
4.1.1. Variáveis	24
Conclusão.....	34
Referências Bibliográficas	35
Apêndice	39

ÍNDICE DE FIGURAS

FIGURA 3.1: Representação gráfico do método de amostragem por validação cruzado de séries temporais.	21
FIGURA 4.1: Evolução do PIBpc para os 51 países.....	23
FIGURA 4.2: Evolução despesas públicas em percentagem do PIB, onde os pontos bordô, azul e verde correspondem respetivamente à educação, saúde e proteção ambiental.	25
FIGURA 5.1: Diagrama de árvore.	30
FIGURA A.1: Correlação das variáveis explicativas.....	42
FIGURA A.2: Evolução da população nos diferentes países.....	43
FIGURA A.3.a: Evolução do índice de direitos de propriedade de Portugal.....	44
FIGURA A.3.b: Evolução do índice de regulamentação de Portugal.....	44
FIGURA A.3.c: Evolução do índice de integridade governamental de Portugal.....	45
FIGURA A.4: Evolução do índice de despesas governamentais de Portugal.....	45
FIGURA A.5.a: Evolução do índice de carga fiscal de Portugal.	46
FIGURA A.5.b: Evolução do índice de saúde fiscal de Portugal.	46
FIGURA A.5.c: Evolução da dívida publica (% PIB) em Portugal.	47
FIGURA A.6.a: Evolução do índice de liberdade de negócio de Portugal.	47
FIGURA A.6.b: Evolução do índice de liberdade laboral de Portugal.	48
FIGURA A.6.c: Evolução da taxa de desemprego em Portugal.	48
FIGURA A.7.a: Evolução do índice de liberdade monetária de Portugal.....	49
FIGURA A.7.b: Evolução da taxa de inflação em Portugal.	49
FIGURA A.8.a: Evolução da balança conta corrente de Portugal.	50
FIGURA A.8.b: Evolução do índice de liberdade comercial de Portugal.....	50
FIGURA A.8.c: Evolução do valor da taxa alfandegária de Portugal.....	51
FIGURA A.9.a: Evolução do índice de liberdade financeira de Portugal.....	51
FIGURA A.9.b: Evolução do índice de liberdade de investimento de Portugal.....	52
FIGURA A.9.c: Evolução do investimento direto estrangeiro (% PIB) em Portugal.	52

FIGURA A.10: Método de PDP aplicado à inflação.	53
FIGURA A.11: <i>Classical feature importance</i> : POLS (esquerda) e Árvore de decisão (direita).	53
FIGURA A.12: <i>Permutation feature importance</i> : POLS (cima (a)), <i>Time-fixed Effects</i> (meio (b)) e XGBoost (baixo (c)).	55
FIGURA A.13: <i>Feature importance aplicado</i> ao modelo XGBoost: frequência (cima (a)), ganho (centro (b)) e cobertura (baixo (c)).	56
FIGURA A.14: Método PDP nos três modelos para as cinco variáveis mais importantes segundo o método SHAP <i>values</i> para XGBoost: PO (esquerda), TFE (centro) e XGB (direita).	58

ÍNDICE DE TABELAS

TABELA 5.1: Hiperparâmetros utilizados para ajustar o modelo XGBoost.	26
TABELA 5.2: Comparação das medidas de precisão preditiva <i>out-of-sample</i> do PIBpc.	27
TABELA 5.3: Comparação das variáveis selecionadas mais importantes nos modelos.	32
TABELA 5.4: Sinal dos efeitos das variáveis mais importantes identificadas pelo SHAP.	33
TABELA A.1: Hiperparâmetros XGBoost.	39
TABELA A.2: Pseudocódigo do método <i>permutation feature importance</i> baseado em Fisher, Rudin, and Dominici (2018).	39
TABELA A.3: Pseudocódigo baseado em Greenwell et al. (2001) para construção da dependência parcial da característica x_1	39
TABELA A.4: Lista de países presentes nos dados de painel para a variável CountryName.	40
TABELA A.5: Lista de variáveis explicativas incluídas nos dados de painel em estudo.	42

AGRADECIMENTOS

Gostaria de agradecer ao orientador Professor João Bastos pelas suas contribuições significativas e pelo apoio quando necessário.

Gostaria de agradecer acima de tudo aos meus pais e à minha irmã pelo apoio constante e paciência ao longo deste período.

Agradeço ainda ao ISEG e a todos os professores e alunos que se cruzaram no meu percurso académico nos últimos anos, em particular às professoras coordenadoras, Professora Margarida Vaz Pato e Professora Cândida Mourão.

Finalmente, um agradecimento particular a Daniela Rodrigues, Iolanda Mota e Andreia Machado pelo apoio, companheirismo e amizade.

1. INTRODUÇÃO

Os últimos 15 anos ficaram marcados por uma profunda instabilidade económica, como elevado desemprego, acentuada oscilação das taxas de juro, baixo crescimento económico, e ainda por novos desafios, tal como o rápido envelhecimento populacional e as alterações climáticas. A economia mundial viu-se ainda abalada por uma pandemia e, mais recentemente, com uma guerra na Europa. É neste contexto de maior volatilidade económica, que modelos complexos de maior precisão produzem vantagens sobre os modelos tradicionais de previsão. Neste trabalho concentramo-nos no crescimento económico, em média por habitante, utilizando a abordagem de dados de painel aplicada a 51 países entre 1995-2022.

O crescimento económico é uma variável macroeconómica traduzida pelo acréscimo de Produto Interno Bruto (PIB) de um país para um determinado período. O maior desafio da estimação do PIB é o número de variáveis a serem tidas em conta. Na literatura, tem vindo a ser propostos vários modelos que serão aprofundados na seção seguinte.

No presente estudo, a variável de previsão é o PIB *per capita* (PIBpc) a preços constantes em dólares americanos, ajustado em paridade de poder de compra. Em termos teóricos, o PIBpc permite quantificar a divisão do produto interno bruto pelo número de habitantes do país, em média. O valor previsto do PIBpc será explicado pela utilização de modelos de Regressão linear, Árvore de decisão e XGBoost para dados de painel. A abordagem de dados de painel permite que muitos dados transversais possam ser observados durante um período definido, enquanto dados transversais inclui muitos objetos em simultâneo.

Segundo a metodologia dos modelos lineares de dados de painel, são propostos vários modelos dos quais serão aprofundados o modelo *Pooled Ordinary Least Squares* (POLS), o modelo de Efeitos Fixos e o modelo de Efeitos Aleatórios. A estes modelos são aplicados testes econométricos para selecionar o modelo mais adequado face às observações. O mesmo será posteriormente utilizado, tal como os restantes modelos, para a comparação de precisão e aplicação de métodos de interpretabilidade.

A interpretabilidade, a par com a explicabilidade, é uma característica fundamental para a confiança no modelo de previsão. Isto significa que quando é possível explicar e compreender internamente um modelo, a previsão obtida pelo mesmo torna-se credível e o modelo utilizável. Nesse sentido, e face ao *trade-off* entre precisão e interpretabilidade, tem vindo a ser desenvolvidos métodos que permitem desmistificar mesmo os modelos mais complexos, em particular modelos *black-box*. O principal objetivo deste estudo é exatamente explorar e aplicar esses métodos a um contexto real.

2. REVISÃO DE LITERATURA

2.1. Modelação da previsão macroeconómica

John M. Keynes, considerado o mais proeminente economista do seu tempo segundo Diebold (1997), publicou, em 1936, o livro *The General Theory of Employment, Interest and Money*. A sua obra impactou para sempre o paradigma da teoria macroeconómica, uma vez que, para além de impulsionar o estudo desta ciência social e criar o conceito de modelo estrutural, também se opôs aos conceitos tradicionais da Economia Clássica. Um modelo estrutural define-se pela relação de interesse entre variáveis explicativas por meio de uma teoria, que incorpora pressupostos sobre especificações do modelo, por exemplo o comportamento do agente económico. Este tipo de modelo permite interpretar dados económicos, e ainda, analisar diferentes cenários prevendo valores futuros para as variáveis explicativas, previsão condicional. Em suma, previsões condicionais exigem modelos estruturais.

O auge do modelo estrutural keynesiano ocorreu na década de 50 e 60 do século XX, devido aos avanços na investigação da econometria macroeconómica estrutural, contudo nas duas décadas seguintes a sua popularidade diminuiu. Para além dos fortes pressupostos assumidos pelo modelo keynesiano, como procura agregada, capacidade produtiva excedentária, rigidez nominal, expectativas adaptativas, entre outros, Lucas (1976) critica a estimativa de parâmetros *ad hoc* e a avaliação quantitativa com base em políticas hipotéticas. Consequentemente, são criadas duas correntes: uma primeira, com o objetivo de aperfeiçoar e corrigir as falhas do sistema tradicional, por exemplo através da integração de expectativas racionais por Fair (1984) e por Taylor (1993); uma outra corrente, com o objetivo de procurar alternativas por meio de modelo não estruturais.

O contínuo progresso sobre o sistema tradicional resultou no surgimento do modelo de Equilíbrio Geral Estocástico Dinâmico (DSGE), uma nova teoria macroeconómica amplamente impulsionada por Lucas (1972) que defendia a integração de processos estocásticos. Este novo modelo consiste numa adaptação da Teoria do Equilíbrio Geral da escola keynesiana, assumindo evolução temporal e princípios microeconómicos, como a teoria de jogos. A vertente mais simples é o modelo linear quadrático que assume agentes com preferências quadráticas, estruturas tecnológicas e funções lineares estocásticas, por exemplo para o consumo e investimento. Define-se, portanto uma nova econometria estrutural, resultante da combinação da teoria macroeconómica moderna e econometria não estrutural de série temporal.

Recuando à década de 20 do século passado, Slutsky (1927) e Yule (1927) apresentaram um quadro de modelação, não estrutural, e previsão de séries cronológicas económicas e financeiras com base em processos autorregressivos. Um modelo de previsão autorregressivo (AR) representa o valor da variável independente, Y_t , é resultado da função linear dos seus valores passados e de

um termo estocástico. Slutsky (1927) e Yule (1927) propuseram ainda um modelo de média móvel, cujo a variável independente é obtida pela função linear de um termo estocástico e de choques aleatórios sucessivos. Contudo, apenas nos anos 70 as previsões econométricas com base em modelos não estruturais cresceram visivelmente, devido ao descontentamento do modelo estrutural, mas fundamentalmente à publicação de um livro de referência de Box e Jenkins (Box et al., 1994). Os autores introduziram o conceito de tendência estocástica, mas a peça central foi o desenvolvimento de modelos lineares de previsão de séries temporais estacionária não sazonal, modelos de média móvel autorregressiva (ARMA), que consistem na combinação simples de modelos autorregressivos e médias móveis. Desenvolveram ainda modelos multivariados, tal como modelos autorregressivos vetoriais (VAR) (Sims, 1980 e Ciccarelli et al., 2016) e modelos autorregressivos vetoriais bayesianos (BVARs) (Carriero et al. 2015), como alternativas menos restritivas a modelos de sistema de equações keynesiano, uma vez que, neste tipo de modelos todas as variáveis são endógenas.

Ao longo dos anos seguintes, destacaram-se várias contribuições ao nível da análise multivariada: Granger (1969) e Sims (1972) defendiam que a noção de causalidade é preditiva, permitindo identificar e compreender ligações cruzadas e feedback de um vetor autorregressivo; também Sargent e Sims (1977) e Geweke (1977) desenvolveram o modelo do fator dinâmico que considera que existem choques económicos idiossincráticos e choques entre sectores, uma vez que os mesmos assumem características e fontes de variação comuns. Esta ideia veio facilitar a modelação e previsão de um grande conjunto de variáveis, em particular no contexto macroeconómico. Modelos não estruturais permitem-nos estimar modelos complicados dispensando a utilização de pressupostos irrealistas, contudo o futuro deste tipo de modelação consiste na utilização combinada de modelos lineares e não lineares, como por exemplo o VAR. Ao modelo VAR são implementadas várias extensões multivariadas do paradigma Box-Jenkins (Diebold, 1998) permitindo aumentar a amplitude da econometria e previsão moderna de séries temporais.

Coulombe et al. (2020) defendem a utilização de modelos não lineares na previsão macroeconómica em contexto de elevada incerteza, stress financeiro, especulação do mercado imobiliário, e acrescido ainda, fenómenos naturais extremos (alterações climáticas), pandemias e guerras. Portanto, num contexto de maior instabilidade e de maior disponibilidade de um grande conjunto de dados macroeconómicos, o *machine learning* tem ganho uma crescente importância para quantificar importantes não-linearidades. A previsão macroeconómica com métodos de *machine learning* remonta ao século anterior, como por exemplo Lee et al. (1993); Kuan e White (1994); Swanson e White (1997); Stock e Watson (1999); Trapletti et al. (2000); Medeiros et al.

(2006). Mais recentemente a capacidade de interpretação dos métodos de *machine learning* tem vindo a ser explorada nas obras publicadas por Joseph (2019) e Zhao e Hastie (2019).

2.2. Indicadores económicos

No caso de estudo, a metodologia adotada é a metodologia do *Index of Economic Freedom*. O índice anual de liberdade económica, disponibilizado pelo *The Heritage Foundation* (2022), é obtido pela pontuação média aritmética, igualmente ponderadas, de doze indicadores, medidos numa escala de 0 a 100, que recaem sobre quatro aspetos fundamentais: Estado de direito, Tamanho do governo, Eficiência regulamentar e Abertura do mercado. Cada um destes componentes será aprofundado em três indicadores que são, em grande medida, calculados com igual ponderação.

O Estado é uma estrutura politicamente organizada, de âmbito político, que se destina a regular vinculativamente o comportamento dos cidadãos e manter a segurança jurídica. Deste modo, o direito de propriedade, a eficiência jurídica e a integridade do próprio governo são aspetos fundamentais de análise governativa. O direito de propriedade diz respeito ao direito do proprietário legal de controlar o acesso a bens legalmente adquiridos pelos mesmos. Este indicador quantifica a eficácia dos direitos de propriedade privada segundo os quadros legais de cada país, com igual foco para os seguintes subfactores: risco expropriação, o respeito dos direitos de propriedade intelectual e a qualidade de execução de contratos e aplicação da lei a este nível. Também a existência de sistema judiciais eficientes e justos são essenciais para garantir a eficácia judicial que por sua vez é medida pela: independência e qualidade do processo judicial, e ainda dos serviços públicos. A corrupção sistémica das instituições governamentais e práticas de suborno, extorsão, nepotismo, compadrio, clientelismo, desvio de fundos e enxerto, constituem as principais motivações de redução da confiança e deterioração da liberdade económica. Neste sentido, este indicador é avaliado equitativamente pela perceção de corrupção no país, controlo da corrupção e risco de suborno.

O governo é a autoridade governamental que coordena a política geral de um país e dirige a Administração pública. A gestão das finanças públicas é, uma das funções do governo, e consiste na captação de recursos e controlo de gastos, garantindo as necessidades coletivas. Neste seguimento, os gastos públicos, a carga fiscal e a própria saúde fiscal são aspetos fundamentais à economia de um país. A metodologia do índice de liberdade económica trata as despesas governamentais de forma não linear, penalizando com maior ênfase, países cujo as despesas governamentais são mais acentuadas, por exemplo economias que excedam 30% do PIB são mais penalizadas do que economias com nível próximo de zero. Neste sentido, países subdesenvolvidos

com pouca capacidade governamental e escassa disponibilização de bens públicos recebem pontuações altas.

A carga fiscal reflete três subfactores quantitativos de igual ponderação: taxa máxima do imposto marginal sobre o rendimento individual, taxa máxima do imposto marginal sobre o rendimento das empresas, e ainda, a carga fiscal total em percentagem do PIB. A saúde fiscal global de um país depende do valor do défice orçamental e do peso da dívida pública no PIB. Se em teoria o financiamento da dívida é visto como uma ferramenta de crescimento económico, o aumento da dívida pública devido a um longo período de défice orçamental, constitui um motivo de estagnação e recessão económica. O cálculo deste indicador tem como ponderação 80% do valor de défices médios em percentagem do PIB nos três anos anteriores e 20% do valor da dívida em percentagem do PIB.

Como já mencionado, o Estado também tem um papel regulamentador e o peso da sua intervenção nos mercados é controverso. A liberdade empresarial, a liberdade laboral e a liberdade monetária são os indicadores considerados para a avaliação da eficiência regulamentar, nomeadamente ao nível dos mercados. A liberdade empresarial pode ser calculada pelo grau de limitação de um país sobre o funcionamento das empresas avaliado sobre os seguintes subfactores: acesso a eletricidade, risco do ambiente de negócios, qualidade regulamentar e inclusão económica das mulheres.

O indicador de liberdade do mercado de trabalho considera a regulamentação e diversos aspetos do quadro legal do mercado de trabalho, destacando o salário mínimo, o direito de associação, as férias anuais pagas, o prazo de pré-aviso para despedimento, a indemnização por despedimento, a produtividade laboral, a taxa de participação da força de trabalho, as restrições ao trabalho suplementar e o despedimento permitido por lei. Todos os subfactores foram avaliados numa escala de 0 a 100, com exceção dos fatores binários, restrições ao trabalho suplementar (sim = 0, não = 100) e despedimento por despedimento permitido por lei (sim = 100, não = 0).

Assumindo como premissa de um mercado livre, a não intervenção do Estado na estabilidade dos preços desse mesmo mercado, a liberdade monetária baseia-se na taxa média ponderada da inflação nos últimos 3 anos, através de uma função convexa, e num critério qualitativo acerca da manipulação governamental sobre os preços, através de imposições diretas e/ou subsídios.

Finalmente, para a análise económica é ainda importante avaliar a abertura da economia aos mercados internacionais, nomeadamente o que diz respeito ao comércio, ao investimento e ao mercado financeiro. A medição da liberdade comercial é composta pela taxa alfandegária média ponderada pelo comércio e pela avaliação qualitativa das barreiras não tarifárias (NTB), uma vez que os dois tipos de barreiras aduaneiras têm impactos na balança corrente dos países. Também

maioria das economias mundiais impõem uma variedade de restrições ao investimento, relativo ao acesso do cambio estrangeiro, pagamentos, transferências de capital, entre outros. Dada essa diversidade de restrições regulamentares, o próprio índice é calculado tendo com base sete aspetos, aos 100 pontos iniciais são deduzidos pontos a cada restrição identificada no país.

O indicador de liberdade financeira permite avaliar a eficiência bancária e o controlo governamental no setor financeiro. Assim sendo, a pontuação atribuída a este indicador tem em consideração cinco critérios: extensão da regulamentação governamental; grau de intervenção do Estado no setor bancário; influência do governo na atribuição de crédito; extensão do mercado financeiro e de capitais; abertura à concorrência estrangeira.

Para além dos indicadores já referidos, a base de dados disponibilizada por *The Heritage Foundation*, contempla outros indicadores económicos e demográficos: população estimada; produto interno bruto, ajustado a paridade do poder de compra, dividido pela população total; taxa de desemprego, força de trabalho que não está empregada mas que procura ativamente trabalho; inflação, variação percentual anual dos preços dos bens e serviços; investimento direto estrangeiro, fluxo total anual de capital próprio, reinvestimento de ganhos, outros capitais de curto e longo prazo, e ainda, empréstimos internacionais de curto e longo prazo; dívida pública bruta em percentagem do PIB; taxa pautal média ponderada pelo comércio.

Em complementaridade ao já proposto, foi introduzido um indicador sobre o comércio internacional, o saldo da balança corrente, que integra a balança de bens e de serviços, através da diferença entre exportações e importações, a balança de rendimentos de fatores produtivos e a balança de transferências unilaterais. Foram ainda introduzidos três indicadores, despesa governamental com a educação, gastos totais em saúde e gastos governamentais com a proteção ambiental, com o objetivo de analisar os impactos económicos do que considero serem três direitos adquiridos e aspetos fundamentais de uma economia sustentável. A integração dos indicadores no modelo é feita sobre o peso dos mesmo no PIB do ano em análise.

2.3. Abordagens à interpretabilidade

Segundo Molnar (2022), *machine learning* consiste num conjunto de métodos computacionais utilizados para melhorar previsões e análises de base de dados. Um modelo de *machine learning* apresenta vantagens ao nível da velocidade, da reprodutibilidade e da escala, contudo este tipo de modelos exige uma volumetria de dados extensa e os insights do processo de aprendizagem são menos perceptíveis, dificultando a interpretação dos mesmos.

Um modelo *black-box* é um modelo cujos mecanismos internos não são revelados, e por isso, o entendimento sobre o seu comportamento e previsão é bastante mais complexo e pouco

compreensível para os humanos, diz-se, portanto, que é um modelo não interpretável. Em oposição, existem modelos *white-box* interpretáveis.

A interpretabilidade é um conceito difícil de definir ao nível matemático. Miller (2017) define a interpretabilidade como o grau em que um humano pode compreender a causa de uma decisão ou previsão. É possível distinguir duas abordagens de interpretabilidade: modelos específicos e modelos agnósticos. Ao contrário dos modelos específicos, modelos agnósticos podem ser aplicados a todos os modelos de *machine learning* e permitem várias formas de explicação e representação. A interpretação através de modelos globais agnósticos é efetuada com base no comportamento médio da amostra de treino (Yang et al., 2018), expressa tendo em conta os valores esperados com base na distribuição dos dados. A interpretação por modelos agnósticos locais significa que a explicação é centralizada numa única observação (Molnar, 2022). De seguida, aprofundarei técnicas integradas nas abordagens mencionadas anteriormente.

O método de substituição Global (Craven e Shavlik, 1996) extrai conclusões do modelo complexo através da interpretação de modelos simples, como Regressão linear ou Árvores de decisão simples. O modelo simples a interpretar corresponde àquele cujo a precisão, medida pelo R-quadrado, se encontra mais próximo de 1, uma vez que o comportamento do modelo interpretável é semelhante ao modelo *black-box*. Este método permite, a todas as pessoas, tirarem conclusões sobre o modelo, contudo o mesmo não tem em consideração os outputs do modelo complexo.

A abordagem da medida de importância, baseado no modelo de Árvores de decisão, mede o impacto na precisão dos modelos devido à exclusão de um determinado regressor, atribuindo pontuações a todos os regressores de um determinado modelo. Quanto maior for a pontuação, maior é o impacto do mesmo no modelo.

A importância do recurso de permutação (Fisher et al., 2018) mede o impacto na precisão dos modelos após a permutação dos valores de um regressor. Este método permite nos compreender o modelo, contudo, tendo em conta que resulta da análise de uma determinada variável/caraterística, não nos permite prever valores para a variável dependente.

O gráfico de dependência parcial representa o efeito marginal que um ou dois regressores têm sobre o resultado preditivo de um modelo de *machine learning* (Friedman, 2001). O gráfico revela a função descrita pela relação de uma variável de resposta e uma variável explicativa, monónica ou mais complexa.

Modelos substitutos locais são modelos interpretáveis que permitem explicar previsões para um subconjunto de dados. Explicações agnósticas de modelos interpretáveis locais (LIME) é uma

técnica proposta por Ribeiro et al. (2016), que explica valores preditivos individuais de forma interpretável e fidedignas. O modelo interpretável a ser utilizado, por exemplo Lasso ou Árvore de decisão, deve ter uma boa aproximação ao modelo local de *machine learning*, não necessariamente uma boa aproximação do modelo global.

Shapley Additive ExPlanations (SHAP) é método de interpretação global, introduzido por Lundberg e Lee (2017), inspirado nos modelos substitutos locais LIME e *Shapley values*. O método *Shapley values* é um outro método de interpretação local inspirado na teoria dos jogos cooperativos, onde cada valor do regressor da instância é um jogador e que a previsão é um pagamento. Deste modo, o output do método de SHAP é uma matriz de valores Shapley, uma linha por instância e uma coluna por característica. A previsão final é obtida pela soma das contribuições que resultam de cada regressor.

3. METODOLOGIA

3.1. Técnicas de modelação

3.1.1. Modelos de interpretabilidade natural

Modelos interpretáveis são modelos cujo processo de análise é compreensível a humanos sem qualquer método de interpretação adicional, tal com Regressão linear, Regressão logística, Árvore de decisão, RuleFit, Naive Bayes, algoritmo dos k - vizinhos mais próximos. Foram selecionados os modelos de Regressão linear para dados de painel e Árvore de decisão.

Modelos de regressão linear

A procura por relações dinâmicas e o controlo da heterogeneidade da seção transversal não observada constituem duas das vantagens de dados das mesmas unidades de seção transversal ao longo do tempo. O modelo de regressão linear é dado por

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + u_{it}, t = 1, 2, \dots, T, \quad (3.1)$$

com uma única observação e um $\boldsymbol{\beta}$ igual em cada período. Para especificar uma determinada i observação de secção transversal utiliza-se y_{it} e as variáveis explicativas \mathbf{x}_{it} que podem variar no tempo ou serem contantes, tal como as variáveis *dummy*. Regra geral, com N grande e T pequeno, as interceções devem ser separadas para cada período, gerando efeitos de tempo agregados que tem a mesma influência no y_{it} para todos os i . Tal que $\boldsymbol{\beta}$ seja consistente para um modelo POLS, assumem-se, de forma suficiente, os seguintes pressupostos

$$E(\mathbf{x}_{it}'u_{it}) = 0, t = 1, 2, \dots, T, \quad (3.2)$$

$$\text{rank}[\sum_{t=1}^T E(\mathbf{x}_t'\mathbf{x}_t)] = K, \text{ para um determinado } i. \quad (3.3)$$

O segundo pressuposto exclui as dependências lineares perfeitas entre variáveis explicativas. De modo a garantir homocedasticidade e nenhuma correlação em série, definiu-se o terceiro pressuposto

$$E(u_t^2 \mathbf{x}_t' \mathbf{x}_t) = \sigma^2 E(\mathbf{x}_t' \mathbf{x}_t) , t = 1, 2, \dots, T , \quad (3.4.a)$$

$$E(u_t u_s \mathbf{x}_t' \mathbf{x}_s) = 0 , t \neq s, t, s = 1, 2, \dots, T , \quad (3.4.b)$$

onde $\sigma^2 = E(u_t^2)$ para todo o t . A parte (a) constitui um pressuposto bastante forte, tal que $E(u_t^2 | \mathbf{x}_t) = E(u_t^2) = \sigma^2$ para todo o t , isto significa que a variância condicional não depende de \mathbf{x}_t e que a variância incondicional é a mesma para cada período. A parte (b) restringe as covariâncias condicionais dos erros a zero, ao longo de dois períodos de tempo diferentes t, s , pelo que implica os pressupostos suficientes $E(u_t u_s) = 0$ e $E(u_t u_s | \mathbf{x}_t, \mathbf{x}_s) = 0, t \neq s, t, s = 1, 2, \dots, T$, e ainda, mais do que uma certa forma de matriz de variância condicional de $\mathbf{u} = (u_1, \dots, u_T)'$. Outro pressuposto consiste em $E(\mathbf{u}_i \mathbf{u}_i') = \sigma^2 \mathbf{I}_T$ que significa que as variações incondicionais são constantes e as covariâncias incondicionais são zero. Verificando-se os pressupostos 1 e 2, pela Propriedade de grandes amostragens de POLS, o estimador OLS agrupado diz-se consistente e assintoticamente normal.

Uma vez que o pressuposto de não correlação entre erro e variáveis explicativas do mesmo período é demasiado forte, tal como é referido por Wooldridge (2002), foram desenvolvidos modelos que incluem efeitos não observados constantes no tempo. O modelo básico de efeitos não observados, para uma observação de corte transversal i é dado por

$$y_{it} = \mathbf{x}_{it} \boldsymbol{\beta} + c_i + u_{it} , t = 1, 2, \dots, T , \quad (3.5)$$

onde \mathbf{x}_{it} é $(1 \times K)$ que pode conter variáveis explicativas observáveis que mudam em t , variáveis explicativas que mudam entre i e variáveis que mudam entre i e t . O efeito individual não observado representado por c_i pode ser ainda designado por variável latente ou heterogeneidade individual não observada. O termo u_{it} representa os erros idiossincráticos ou distúrbios idiossincráticos porque variam em t e i . Por uma questão de simplificação, a equação pode ser escrita da seguinte forma

$$y_{it} = \mathbf{x}_{it} \boldsymbol{\beta} + v_{it} , t = 1, 2, \dots, T , \quad (3.6)$$

onde $v_{it} = c_i + u_{it} , t = 1, 2, \dots, T$, designa-se por erros compostos. Para cada t , v_{it} é a soma do efeito individual não observado e o erro idiossincrático. Pelo mesmo raciocínio do modelo linear simples, o estimador POLS diz-se consistente se $E(\mathbf{x}_{it}' v_{it}) = 0, t = 1, 2, \dots, T$. Assume-se os seguintes pressupostos para não correlação entre \mathbf{x}_{it} e v_{it}

$$E(\mathbf{x}_{it}' u_{it}) = 0 \text{ e } E(\mathbf{x}_{it}' c_i) = 0 , t = 1, 2, \dots, T , \quad (3.7)$$

Se $E(y_{it} | \mathbf{x}_{it}, c_i) = 0$ for bem definido então $E(\mathbf{x}_{it}' u_{it}) = 0$. Modelos com variáveis dependentes desfasadas em \mathbf{x}_{it} devem violar os pressupostos acima, uma vez que inevitavelmente $y_{i,t-1}$ e c_i são correlacionados.

Apesar dos pressupostos (3.7), os erros compostos v_{it} serão correlacionados em série devido a c_i em cada t : v_{it} não diminui com o aumento de $|t - s|$, logo v_{it} não revela uma fraca dependência ao longo do tempo. Deste modo e considerando que para cada (y_i, \mathbf{X}_i) existem T linhas ordenadas cronologicamente e empilhadas de $i = 1, 2, \dots, N$, é importante definir um N grande e T assintóticos fixos para um bom desempenho do método POLS. A inferência POLS requer um estimador de matriz de variância robusto e testes estatísticos robustos.

O modelo de efeitos aleatórios (RE) integra c_i no termo do erro composto, tal como acontece com o modelo POLS, contudo impõe mais suposições

$$E(u_{it} | \mathbf{x}_i, c_i) = 0, \quad t = 1, 2, \dots, T, \quad (3.8.a)$$

$$E(c_i | \mathbf{x}_i) = E(c_i) = 0, \quad \text{onde } \mathbf{x}_i = \{\mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots, \mathbf{x}_{iT}\}. \quad (3.8.b)$$

A equação (3.8.a) representa o pressuposto de exogeneidade rigorosa. A equação (3.8.b) define a ortogonalidade entre c_i e cada \mathbf{x}_{it} . De modo a obter resultados consistentes o pressuposto (3.8.b) pode ser relaxado, contudo este é implícito pelo facto de, x_{it} ser fixo e $E(c_i) = 0$, ou pelo pressuposto de que c_i é independente de \mathbf{x}_i .

A abordagem de efeitos aleatórios explora a correlação em série no erro composto v_{it} através do método dos mínimos quadrados generalizados (GLS). O pressuposto de exogeneidade para GLS é dada por $E(v_{it} | \mathbf{x}_i) = 0, \quad t = 1, 2, \dots, T$. Se satisfaz o pressuposto, podemos utilizar o método GLS com $v_{it} = c_i + u_{it}$. Para todos os períodos em T , podemos escrever

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{v}_i, \quad \text{onde } \mathbf{v}_i = c_i \mathbf{j}_T + \mathbf{u}_i \text{ e } \mathbf{j}_T \text{ é o } (T \times 1) \text{ vetor.} \quad (3.9)$$

A matriz de variância incondicional de \mathbf{v}_i é definido por $\boldsymbol{\Omega} = E(\mathbf{v}_i \mathbf{v}_i')$. Uma matriz $T \times T$ necessariamente igual para todos os i assumindo a amostragem aleatório em *cross section*. De modo a obter um valor consistente de GLS é também necessário satisfazer o segundo pressuposto do modelo de efeitos aleatórios, *rank condition*, para GLS:

$$\text{rank}[\sum_{t=1}^T E(\mathbf{X}_t' \boldsymbol{\Omega}^{-1} \mathbf{X}_t)] = K. \quad (3.10)$$

A análise de efeitos aleatórios adiciona o pressuposto que o erro idiossincrático é dado sobre a forma de $\boldsymbol{\Omega}$. A primeira condição consiste em os erros idiossincráticos terem uma variação incondicional constante ao longo do t , e o segundo pressuposto é que os erros idiossincráticos são em série não correlacionados. Os dois pressupostos são definidos por

$$E(u_{it}^2) = \sigma_u^2, \quad t = 1, 2, \dots, T, \quad (3.11.a)$$

$$E(u_{it}u_{is}) = 0, \quad t \neq s, t, s = 1, 2, \dots, T. \quad (3.11.b)$$

Sobre a hipótese de $E(v_{it}^2) = E(c_i^2) + 2E(c_i u_{it}) + E(u_{it}^2) = \sigma_c^2 + \sigma_u^2$, onde $\sigma_c^2 = E(c_i^2)$.
 $E(v_{it}v_{is}) = E[(c_i - u_{it})(c_i - u_{is})] = E(c_i^2) = \sigma_c^2$, para todo o $t \neq s$. (3.12)

O terceiro pressuposto consiste em

$$E(\mathbf{u}_i \mathbf{u}_i' | \mathbf{x}_i, c_i) = \sigma_u^2 \mathbf{I}_T, \quad (3.13.a)$$

$$E(c_i^2 | \mathbf{x}_i) = \sigma_c^2. \quad (3.13.b)$$

Sob o pressuposto acima, $E(u_{it}^2 | \mathbf{x}_i, c_i) = \sigma_u^2$, $t = 1, 2, \dots, T$ que é implícito por (10.28) e $E(u_{it}u_{is} | \mathbf{x}_i, c_i) = 0$, $t \neq s, t, s = 1, 2, \dots, T$. O terceiro pressuposto assume que as variações condicionais são constantes e as covariâncias condicionais são igual a zero. A junção do pressuposto (3.11.b) e (3.13.b) resulta numa hipótese de homocedasticidade sobre o efeito não observado c_i , $Var(c_i | \mathbf{x}_i) = Var(c_i)$.

Para implementação do procedimento FGLS, definido por $\sigma_v^2 = \sigma_c^2 + \sigma_u^2$. Assumindo o estimador consistente de σ_u^2 e σ_c^2 , então $\hat{\Omega} = \hat{\sigma}_u^2 \mathbf{I}_T + \hat{\sigma}_c^2 \mathbf{j}_T \mathbf{j}_T'$. O estimador FGLS com efeitos aleatórios é dado por

$$\hat{\beta}_{RE} = (\sum_{i=1}^N \mathbf{X}_i' \hat{\Omega}^{-1} \mathbf{X}_i)^{-1} (\sum_{i=1}^N \mathbf{X}_i' \hat{\Omega}^{-1} \mathbf{y}_i). \quad (3.14)$$

O estimador de efeitos aleatórios é claramente motivado pelo terceiro pressuposto, porém $\hat{\beta}_{RE}$ é sempre consistente. Se o primeiro pressuposto e a *rank condition* se mantenha, então $\hat{\beta}_{RE} \xrightarrow{p} \beta$ com $N \rightarrow \infty$.

Em alternativa aos modelos de efeitos aleatórios, foram desenvolvidos modelos que admitem correlação arbitrária entre os dados de painel e \mathbf{x}_{it} , modelos de efeitos fixos (FE), destacados na investigação de Simionescu et al. (2016). O primeiro pressuposto é a exogeneidade estrita das variáveis explicativas condicionadas a c_i :

$$E(u_{it} | \mathbf{x}_i, c_i) = 0, \quad t = 1, 2, \dots, T. \quad (3.15)$$

Este pressuposto é idêntico ao primeiro pressuposto de RE, contudo a diferença prende-se na não definição de (3.13.b), uma vez que FE define que $E(c_i | \mathbf{x}_i)$ pode ser qualquer função de \mathbf{x}_i . A análise de efeitos fixos é mais robusta comparativamente, não sendo possível distinguir efeitos observáveis constantes no tempo e efeitos não observados constantes no tempo, uma vez que c_i pode ser correlacionado arbitrariamente com cada elemento \mathbf{x}_{it} . Por consequência, e não incluindo mais pressupostos, não é possível incluir fatores de tempo em \mathbf{x}_{it} . Quando analisando

características qualitativas individuais, nomeadamente género, etnia ou nacionalidade, estas características podem ser incluídas em \mathbf{x}_{it} . Wooldridge (2002) explica que a indústria e variáveis que descrevem atributos fixos de uma cidade não podem ser incluídas em \mathbf{x}_{it} , a menos que pelo menos uma empresa ou cidade sofra uma alteração ao longo do tempo. Cada elemento \mathbf{x}_{it} varia ao longo do tempo para algumas unidades de secção transversal, designamos por “variáveis explicativas variáveis no tempo”.

Em termos gerais, seja $d2_t, \dots, dT_t$ as variáveis *dummies* no período t , então $ds_t = 1$ se $s = t$ e $ds_t = 0$ caso contrário. Seja \mathbf{z}_i o vetor de variáveis observáveis constantes no tempo, e \mathbf{w}_{it} o vetor de variáveis, variáveis no tempo, então y_{it} é dado por

$$y_{it} = \theta_1 + \theta_2 d2_t + \dots + \theta_T dT_t + \mathbf{z}_i \gamma_1 + d2_t \mathbf{z}_i \gamma_2 + \dots + dT_t \mathbf{z}_i \gamma_T + \mathbf{w}_{it} \boldsymbol{\delta} + c_i + u_{it}, \quad (3.16)$$

$$E(u_{it} | \mathbf{z}_i, \mathbf{w}_{i1}, \mathbf{w}_{i2}, \dots, \mathbf{w}_{iT}, c_i) = 0, \quad t = 1, 2, \dots, T. \quad (3.17)$$

A equação (3.17) representa a relação causal pretendida, no qual se condiciona c_i para controlar os fatores não observáveis constantes no tempo. Sem mais pressupostos, a intercepção θ_1 pode não ser identificada e o vetor γ_1 em \mathbf{z}_i pode ser identificado porque $\theta_1 + \mathbf{z}_i \gamma_1$ não pode ser distinguido de c_i , onde θ_1 é a intercepção para o período de tempo $t = 1$ e γ_1 mede os efeitos de \mathbf{z}_i em y_{it} no período $t = 1$. Apesar de não conseguirmos identificar os efeitos \mathbf{z}_i , conseguimos identificar $\gamma_1, \gamma_2, \dots, \gamma_T$ e desse modo podemos estimar as diferenças nos efeitos parciais nas variáveis temporais para um determinado período e efeitos das variáveis temporais se variam no tempo.

O estimador de efeitos fixos $\boldsymbol{\beta}$ sob o primeiro pressuposto (3.15) é obtido da transformação da equação com o objetivo de eliminar o efeito não observado c_i . A transformação de efeitos fixos, também designada por *within transformation*, é dado por

$$\bar{y}_i = \bar{\mathbf{x}}_i \boldsymbol{\beta} + c_i + \bar{u}_i, \quad (3.18)$$

onde $\bar{y}_i = T^{-1} \sum_{t=1}^T y_{it}$, $\bar{\mathbf{x}}_i = T^{-1} \sum_{t=1}^T \mathbf{x}_{it}$, e $\bar{u}_i = T^{-1} \sum_{t=1}^T u_{it}$. Para cada t , a equação de efeitos fixos transformados é dada por

$$\dot{y}_{it} = \dot{\mathbf{x}}_{it} \boldsymbol{\beta} + \dot{u}_{it} \Leftrightarrow y_{it} - \bar{y}_i = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i) \boldsymbol{\beta} + (u_{it} - \bar{u}_i), \quad t = 1, 2, \dots, T. \quad (3.19)$$

A relaxação temporal traduz-se na eliminação do efeito individual fixo c_i . A interpretação de $\boldsymbol{\beta}$ deriva da expectativa condicional $E(y_{it} | \mathbf{x}_i, c_i) = E(y_{it} | \mathbf{x}_{it}, c_i) = \bar{\mathbf{x}}_{it} \boldsymbol{\beta} + c_i$. Para que a estimativa POLS seja consistente, tem de se garantir a hipótese OLS

$$E(\dot{\mathbf{x}}_{it}' \dot{u}_{it}) = 0, \quad t = 1, 2, \dots, T. \quad (3.20)$$

Sobre o pressuposto FE.1, u_{it} e \bar{u}_i não revelam correlação com \mathbf{x}_{it} e $\bar{\mathbf{x}}_i$ para todo o $t = 1, 2, \dots, T$, então podemos assumir que (3.20) também produz estimadores OLS consistentes. É ainda possível assumir que $E(\ddot{u}_{it}|\mathbf{x}_i) = E(u_{it}|\mathbf{x}_i) - E(\bar{u}_i|\mathbf{x}_i) = 0$ que por sua vez implica que $E(\ddot{u}_{it}|\ddot{\mathbf{x}}_{i1}, \ddot{\mathbf{x}}_{i2}, \dots, \ddot{\mathbf{x}}_{iT}) = 0$, tal que $\ddot{\mathbf{x}}_{it}$ é apenas uma função de $\mathbf{x}_i = (\mathbf{x}_{i1}, \dots, \mathbf{x}_{iT})$ e satisfaz a forma de expectativa condicional do pressuposto de exogeneidade estrita na equação (3.20). O estimador de efeitos fixos representado por $\hat{\beta}_{FE}$ é o estimador POLS para a regressão \ddot{y}_{it} em $\ddot{\mathbf{x}}_{it}$, $t = 1, 2, \dots, T$, $i = 1, 2, \dots, N$.

O segundo pressuposto de FE define que, para que o estimador tenha um comportamento assintoticamente bem-comportado, a *rank condition* normalizada na matriz de variáveis explicativas *time-demeaned* é

$$\text{rank}[\sum_{t=1}^T E(\ddot{\mathbf{x}}_{it}' \ddot{\mathbf{x}}_{it})] = \text{rank}[(\ddot{\mathbf{X}}_i' \ddot{\mathbf{X}}_i)] = K. \quad (3.21)$$

A equação (3.21) explica que, se \mathbf{x}_{it} inclui um elemento que não varia ao longo do tempo para qualquer i , então $\ddot{\mathbf{x}}_{it}$ é igual a zero para todos os t , e conseqüentemente $\ddot{\mathbf{X}}_i$ incluiria uma coluna de zeros para os respectivos i .

Com o objetivo de assegurar que o estimador FE é sempre consistente define-se o terceiro pressuposto, idêntico ao pressuposto da equação (3.13.a), dado por

$$E(\mathbf{u}_i \mathbf{u}_i' | \mathbf{x}_i, c_i) = \sigma_u^2 \mathbf{I}_T. \quad (3.22)$$

Seja $E(\mathbf{u}_i | \mathbf{x}_i, c_i) = 0$ pelo primeiro pressuposto, então o terceiro pressuposto assume que $\text{Var}(\mathbf{u}_i | \mathbf{x}_i, c_i) = \sigma_u^2 \mathbf{I}_T$. O estimador de efeitos fixos, também designado por *within estimator*, é expresso por

$$\hat{\beta}_{FE} = (\sum_{i=1}^N \ddot{\mathbf{X}}_i' \ddot{\mathbf{X}}_i)^{-1} (\sum_{i=1}^N \ddot{\mathbf{X}}_i' \ddot{\mathbf{y}}_i) = (\sum_{i=1}^N \sum_{t=1}^T \ddot{\mathbf{x}}_{it}' \ddot{\mathbf{x}}_{it})^{-1} (\sum_{i=1}^N \sum_{t=1}^T \ddot{\mathbf{x}}_{it}' \ddot{\mathbf{y}}_{it}). \quad (3.23)$$

O estimador utiliza a variação temporal dentro de cada seção transversal.

Árvore de decisão

O modelo de Regressão de árvore é um modelo não linear e não paramétrico, construído sobre a estrutura de uma árvore. O nó raiz contém todas as observações, e através do algoritmo *greedy search*, o conjunto de dados é recursivamente dividido em conjuntos de dados de menor dimensão. Em cada nó de decisão, são gerados diferentes ramos que representam subconjuntos distintos. Os subconjuntos finais não subdivididos dizem-se nós folha. O objetivo é minimizar a variância da variável dependente, garantido que o *outcome* no nó filho é sempre mais homogênea do que o nó pai. A interpretação dos algoritmos de indução de árvore é semelhante para a maioria dos tipos de árvore. Deste modo, relação entre o resultado y e os regressores x , é dado por:

$$\hat{y} = \hat{f}(x) = \sum_{m=1}^M c_m I_{\{x \in R_m\}}, \quad (3.24)$$

onde cada instância corresponde a um subconjunto R_m , nó folha. $I_{\{x \in R_m\}}$ é uma função identidade que retorna 1 se x se encontrar no subconjunto R_m , 0 caso contrário. A redução da variância da variável dependente é medida pelo ‘*sum of squares reduction*’ dado por

$$SSRe = SST - (SSL + SSR), \quad (3.25)$$

onde $SST = \sum_{i=1}^n (y_i - \bar{y})^2$ é a soma dos quadrados totais para o nó de decisão, SSL a soma ao quadrado do nó esquerdo e SSR a soma ao quadrado do nó direito. A previsão é limitada entre 0 e 1, na medida em que os valores previstos são iguais à média da variável dependente.

Após o modelo estar completo, é possível aplicar uma técnica que permite reduzir a dimensão da árvore. O *pruning* é particularmente importante em árvores complexas, pois remove nós redundantes e não críticos aos dados. Deste modo, esta técnica melhora a precisão do modelo e reduz o risco de sobre ajustamento dos dados, *overfitting*.

A explicabilidade do modelo é principal vantagem do modelo de regressão de árvore. Para além da sua representação natural, a sua estrutura permite a análise de cenários hipotéticos. É ainda possível utilizar o método de importância dos regressores, onde cada importância pode ser interpretada como parte da importância geral do modelo, ou ainda o método de decomposição da árvore, que permite rastrear uma decisão e explicar a previsão pelas contribuições adicionadas em cada nó de decisão. A previsão final de uma instância, segundo a decomposição da árvore (Molnar, 2022), é obtida através da seguinte equação:

$$\hat{f}(x) = \bar{y} + \sum_{d=1}^D split * contribuição_{(d,x)} = \bar{y} + \sum_{j=1}^p regressor * contribuição_{(j,x)}. \quad (3.26)$$

A previsão corresponde à soma do valor médio do *outcome* e do somatório das contribuições para os D ramos entre o nó raiz e o nó terminal. Podemos ainda interpretar o impacto de cada uma das p características na previsão.

3.1.2. Black-box model: eXtreme Gradient Boosting (XGBoost)

XGBoost é um método de *machine learning*, definido por Chen e Guestrin (2016), como um sistema escalável de reforço de árvore (*tree boosting*). Isto significa que é um método baseado em Árvores de decisão com métodos de regularização do aumento do gradiente (*gradient boosting*), garantindo uma elevada precisão, eficiência e flexibilidade.

Gradient Descent

O algoritmo de descida do gradiente tem como objetivo a minimização da função de perda. Para um par de observações $z = (x, y)$ e função $F_w(x)$, onde w é um parâmetro que mapeia as variáveis explicativas x para a observação y , obtemos a função de perda diferenciável $L(z, w)$. O objetivo é utilizar o gradiente $\nabla_w L(z, w)$, respeitando os diferentes parâmetros e pesos da função, de modo a minimizar a função de perda. O primeiro passo consiste em definir o valor inicial w_0 . Seguidamente o valor do gradiente é calculado com base em cada par de observações z_i . Os valores seguintes são calculados da seguinte maneira

$$w_{t+1} = w_t - \gamma \frac{1}{n} \sum_{i=1}^n (\nabla_w L(z_i, w_t)), \quad (3.27)$$

onde w_{t+1} é o parâmetro ótimo utilizando a derivada da função de perda $\nabla_w L(z_i, w_t)$, para um par de observações $z_i = (x_i, y_i)$. O parâmetro γ é a taxa de aprendizagem que se mantém sempre positivo. Quando esta se encontra num valor suficientemente pequeno, o algoritmo converge para um mínimo local. Seja ϵ um valor suficientemente pequeno, o critério de paragem é dado por $\|F(x_n)\| < \epsilon$.

Boosting

Boosting é definido por Kearns (1988) e Schapire (1990) como um conjunto combinado de Árvores decisão processado de forma sequencial, a fim de obter um menor erro da previsão. A seleção do subconjunto de dados de treino utilizado em cada membro é efetuada com base no desempenho do modelo anterior. Os pesos maiores são atribuídos às observações incorretas do modelo anterior e, por isso, estas são escolhidas para integrar o próximo membro com mais frequência do que as observações corretas. Deste modo, após vários ciclos, o método de *boosting* combina observações fracas numa única observação.

Para um dado conjunto de dados com n observações e m regressores, $D = (x_i, y_i) (|D| = n, x_i \in \mathbb{R}^m, y_i \in \mathbb{R})$, D corresponde à equação de um modelo de conjunto de árvores com K funções aditivas para previsão do output

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i), f_k \in F. \quad (3.28)$$

O espaço da regressão de árvore é dado por $F = f(x) = w_{q(x)} (q: \mathbb{R}^m \rightarrow T, w \in \mathbb{R}^T)$, onde q representa a estrutura de cada árvore que mapeia uma observação para o índice de folha correspondente e T é o número de folhas numa árvore. A cada f_k corresponde uma estrutura de árvore independente q e os pesos da folha w .

Algoritmo XGBoost

A função de perda usada pelo algoritmo XGBoost corresponde à combinação do erro quadrático médio da função de perda e de um termo de penalidade Ω . O erro quadrático médio calcula o erro do PIBpc prevista, \hat{y}_1 , utilizando o PIBpc real. Uma vez que a função inclui um termo de penalidade, protegendo contra o *overfitting*, a função de perda pode ser chamada de função de perda regularizada:

$$L(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k), \quad (3.29)$$

com $\Omega(f) = \gamma T + \frac{1}{2} \lambda \|w\|^2$ e $l(\hat{y}_i, y_i) = (y_i - \hat{y}_i)^2$. O parâmetro T é o número de folhas numa árvore, γ e λ são parâmetros de regularização, f_k corresponde à estrutura da árvore independente q e os pesos da folha w , e l a função de perda convexa diferenciável dada pela média quadrática do erro.

A função de perda reduz quando é adicionada uma nova árvore, tal que a complexidade do modelo a adicionar não excede o valor acrescentado do novo modelo de árvore, e, portanto, os valores para o PIBpc são melhor modeladas. \bar{y}_i^t é o PIBpc previsto da t -ésima interação da i -ésima instância. O objetivo é adicionar a f_t árvore que minimiza a função de perda:

$$L^t = \sum_{i=1}^n l(y_i + \hat{y}_i^{t-1} + f_t(x_i)) + \Omega(f_t), \quad (3.30)$$

f_t , que melhora o modelo 3.8., adicionado segundo o algoritmo *greedy search*. XGBoost utiliza a aproximação de segunda ordem para otimizar a equação 3.29.

$$L^t \approx \sum_{i=1}^n [l(y_i + \hat{y}_i^{t-1}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t), \quad (3.31)$$

onde $g_i = \frac{\delta}{\delta \hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$ é a primeira derivada da função de perda l , $h_i = \frac{\delta^2}{\delta^2 \hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$ é a segunda derivada da função de perda l e \hat{y}_i^{t-1} o PIBpc prevista da iteração anterior. O objetivo é minimizar a função perda da regressão de árvore f_t para cada iteração t . Por este motivo, termos que não dependam da nova árvore f_t podem ser removidos sem qualquer impacto da função objetivo

$$\tilde{L} = \sum_{i=1}^n [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t). \quad (3.32)$$

O algoritmo XGBoost utiliza o método de Newton *boosting*, que diferentemente do *gradient boosting* introduzido por Friedman (2001), baseia-se na atualização do gradiente descendente de segunda ordem. Este fator, tal como Sigrist (2018) demonstra, apresenta uma melhor performance face a outros métodos de *boosting*.

O segundo aspeto relevante do algoritmo consiste em encontrar os *split points*, os pontos de separação, que dividem entre folhas e ramos. Definindo $I_j = i | f_t(x_i) = j$ como o conjunto de observações i na folha ou nó terminal j da árvore t . À função objetivo anterior aprofundamos o termo Ω , somando as instâncias de cada folha $j, i \in I_j$, tal como apresentado de seguida.

$$\tilde{L}^t = \sum_{i=1}^n \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 = \sum_{j=1}^T \left[\left(\sum_{i \in I_j} g_i \right) w_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) w_j^2 \right] + \gamma T. \quad (3.33)$$

A última parte da expressão matemática pretende separar todas as observações i em conjuntos de observações I_j , tal que o valor para $f_t(x_i)$ em conjuntos de observação I_j é obtido por w_j , e as derivadas g_i e h_i , que dependem da função de perda da previsão anterior, tornam-se constantes. Para uma estrutura de árvore fixa $q(x)$ com conjunto de observações I_j em cada folha j , o valor ótimo para o peso w_j^* é o valor ótimo do PIBpc, que é obtido pela equação quadrática

$$w_j^* = - \frac{\sum_{i \in I_j} g_i}{\left(\sum_{i \in I_j} h_i + \lambda \right)}. \quad (3.34)$$

Por substituição, o valor mínimo da função perda de uma determinada partição é dado por

$$\tilde{L}^t(f_t) = - \frac{1}{2} \sum_{j=1}^T \frac{\left(\sum_{i \in I_j} g_i \right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T. \quad (3.35)$$

Assumindo que L_L e L_R são o conjunto de instâncias dos nós da esquerda e direita, respetivamente, então $I = I_L \cup I_R$. Após a divisão por ramos, a função perda é

$$L_{split} = \frac{1}{2} \left[\frac{\left(\sum_{i \in I_L} g_i \right)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{\left(\sum_{i \in I_R} g_i \right)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{\left(\sum_{i \in I} g_i \right)^2}{\sum_{i \in I} h_i + \lambda} \right] - \gamma. \quad (3.36)$$

Para obter a *score* de qualidade da previsão (Chen e Guestrin, 2016), é necessário apenas atribuir valores ao gradiente e ao gradiente de segunda ordem e efetuar as operações.

Otimização de hiperparâmetros

Um hiperparâmetro é uma característica externa a um modelo, cujo a definição do seu valor é feita numa fase inicial do processo de aprendizagem, uma vez que não pode ser estimado a partir dos dados. Portanto, uma das principais etapas da modelagem em XGBoost consiste na definição de hiperparâmetros, tal como a taxa de aprendizagem, peso mínimo da criança, profundidade máxima da árvore, percentagem da amostra da coluna, percentagem de subamostra e *gamma*. A Tabela A.1. em Apêndice propõe valores pré-especificados para cada um dos hiperparâmetros referidos, em semelhança ao apresentado por Matos (2020). *Grid Search* é um algoritmo de pesquisa simples, que recebe um conjunto de valores para um ou mais hiperparâmetros e testa

todas as combinações possíveis dentro dessa vizinhança de modo a aproximar-se do objetivo, encontrar os valores ótimos para os hiperparâmetros.

3.2. Abordagens de interpretabilidade selecionadas

3.2.1. Classical Feature Importance

Para avaliação de modelos de árvore são considerados três indicadores clássicos de importância da característica, *feature importance*: cobertura, frequência e ganho. A cobertura é expressa, em percentagem, como as métricas de cobertura dos regressores. A cobertura para cada regressor consiste no número de observações obtidas pela soma entre o número de pontos atribuídos a cada divisão de dados e o número de observações, que o próprio regressor, é utilizado para dividir os dados sobre todas as árvores. Por sua vez, a frequência corresponde ao número total de vezes que um regressor é utilizado para dividir os dados entre todas as árvores. A frequência para cada regressor corresponde, em percentagem, ao peso da frequência do mesmo regressor sobre a frequência de todos os regressores. Finalmente, o ganho tem em consideração a redução média na perda de treino alcançada utilizando uma característica dividida. Quanto maior o valor de ganho, maior a importância do regressor na previsão.

3.2.2. Permutation Feature Importance

A abordagem permutação da importância da característica, proposto por Fisher et al. (2018), mede a variação do erro de previsão do modelo após a permutação dos valores dos regressores, o que quebra a relação entre o regressor e o verdadeiro output. O regressor diz-se “importante” quando existe um aumento do erro após a alteração do valor do regressor. A função de perda é calculada pelo erro médio absoluto (EAM) obtido através de

$$EAM = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|. \quad (3.37)$$

Na Tabela A.2. em Apêndice, o pseudocódigo do algoritmo de *permutation feature importance*.

3.2.3. Partial Dependence Plots (PDP)

O gráfico de dependência parcial (PDP), introduzido por Friedman (2001), permite mostrar os efeitos marginais de um regressor, normalmente, na previsão de um modelo. Esta abordagem permite identificar a relação entre o regressor e o *outcome*.

Tal como demonstrado por Friedman (2001) e Greenwell et al. (2001), seja $\hat{f}(x)$ a função de previsão, onde $x = x_1, x_2, \dots, x_p$ representa as características do modelo e z_s é o conjunto de interesse particionado por x , então a "dependência parcial" de z_s é definido por

$$f_s(z_s) = E_{z_c}[\hat{f}(z_s, z_c)] = \int \hat{f}(z_s, z_c) p_c(z_c) dz_c, \quad (3.38)$$

onde $z_c = x / z_s$ e, $p_c(z_c)$ a densidade da probabilidade marginal de z_c , dado por $p_c(z_s) = \int p(x) dz_s$. Aplicando a equação (3.37) para amostra de treino, obtemos a estimativa

$$\bar{f}_s(z_s) = \frac{1}{n} \sum_{i=1}^n \hat{f}(z_s, z_{i,c}), \quad (3.39)$$

onde $z_{i,c}$, ($i = 1, 2, 3, \dots, n$) são os valores de z_c para a amostra de treino.

O gráfico de dependência parcial para x_1 , seja $z_s = x_1$ a variável preditora de interesse com vetor $x_{11}, x_{12}, \dots, x_{1k}$, pode ser construído segundo o pseudocódigo da Tabela A.3. em Apêndice.

Apesar do gráfico de dependência parcial ser fácil de implementar e interpretar, por vezes o pressuposto de independência pode constituir um problema, uma vez que, quando existe alta correlação entre regressores, o valor de output pode ser enviesado e levar a interpretações erradas. Também a ocultação dos efeitos heterogêneos na estimativa de PDP pode levar a conclusões erradas, neste caso sobre a característica de impacto na previsão.

3.2.4. SHAP (SHapley Additive ExPlanations) Values

Additive feature attribution methods

SHAP é um modelo agnóstico que pertence a uma classe denominada por “métodos de atribuição de características aditivas”, uma vez que a explicação é dada como uma função linear de características. A explicação do modelo f é dado pela soma de um termo de interceção ϕ_0 e de output z' com o valor atribuído para cada característica ϕ_t .

$$f(z') = \phi_0 + \sum_{t=1}^M \phi_t z'_t, \quad (3.40)$$

onde $\phi_t \in IR$, $z' \in \{0, 1\}$, M é o número de características interpretáveis e z'_t corresponde ao aparecimento das características. SHAP substitui cada característica x_i por uma variável binária z'_t , no qual $z'_t = 1$, se a característica foi observada e $z'_t = 0$, caso contrário.

Shapley Values

O método SHAP tem como base os valores de Shapley (Shapley, 1953) obtidos da teoria dos jogos. Neste contexto, um jogo simples, com uma única ronda, corresponde à tarefa de prever para uma única observação de conjunto de dados. O ganho do jogador corresponde ao valor estimado menos a previsão média para todas as observações. Os jogadores são os valores dos regressores para as observações que recebem ganho.

O acréscimo do valor de uma característica devido à variação ponderada do valor *outcome* do modelo pode ser calculado quando adicionamos uma característica i em todos os subconjuntos de características que excluem a mesma, isto é $(S \subseteq F \setminus \{i\})$. O valor SHAP para características num modelo que indica o atributo das características para a previsão do *outcome*. O valor SHAP para a característica i e modelo f é dado por

$$\phi_i(f) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} \left(f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right), \quad (3.41)$$

Onde F é o conjunto total de características para o modelo f , $(S \subseteq F \setminus \{i\})$ são todos os possíveis subconjuntos de F excluindo a característica i , $f_S(x_S)$ é a função treinada pelas características do subconjunto S e $f_{S \cup \{i\}}(x_{S \cup \{i\}})$ é a função treinada pelas características do subconjunto $S \cup \{i\}$.

SHAP Feature Importance

O SHAP considera como mais característica mais importante, a característica com maior valor Shapley absoluto. Contudo o objetivo é medir para cada característica a importância global dada pela média dos valores Shapley absolutos para cada característica de um conjunto de dados.

$$I_i = \sum_{j=1}^n |\phi_i^j|, \quad (3.42)$$

onde i representa uma característica e j uma observação do conjunto de dados.

O gráfico de importância da característica representa as variáveis de ordem decrescente de importância. O método de importância da característica pelos valores de Shapley baseia-se nas contribuições do regressor, enquanto outros métodos, como por exemplo o método de permutação da importância da característica que se baseia na precisão do modelo.

3.3. Amostragem de treino e teste

Com o objetivo de avaliar a previsão, os dados são divididos em amostra de treino e amostra de teste. O modelo é estimado com base na amostra de treino e, posteriormente, a precisão do mesmo é avaliada utilizando medidas de erro aplicadas à amostra de teste. Quando este procedimento é efetuado apenas uma vez, a avaliação diz-se *fixed origin*. Porém, devido à existência de *outliers* ou mudanças de nível, o método mais adequado poderá consistir num método de avaliação no qual a origem da amostra de teste é atualizada sucessivamente e as previsões são produzidas a partir de uma nova origem, *rolling origin* (Tashman, 2000).

O método de reamostragem por validação cruzada para séries temporais é um método de reamostragem não aleatório, cujo conjunto de dados são ordenados por ordem cronológica. A amostra de teste é constituída por h observações, que corresponde aos períodos utilizados na avaliação preditiva do modelo. A amostra de treino corresponde às observações anteriores às observações de teste e consiste no conjunto de observações para estimação do modelo. O plano de validação cruzada de séries temporais distingue-se dos restantes métodos, pelo facto de começar com as observações de séries temporais mais recentes e recursivamente retroceder no tempo. Isto significa que a primeira reamostragem é obtida pela divisão total dos dados.

A Figura 3.1. abaixo retrata a ideia básica da amostragem por *rolling origin cross validation* para séries temporais. As observações vermelhas representam as amostras de teste, que corresponde a 20% do número total dos anos em estudo, aproximadamente 5 anos. Os dados de painel utilizados no presente estudo garantem a existência do mesmo conjunto de países nos diferentes períodos em análise. As observações azuis representam as amostras de treino que correspondem às observações dos períodos anteriores à amostra de teste. As previsões são produzidas a partir de iterações para 6 origens distintas, no qual a primeira iteração corresponde às observações mais recentes. O modelo é reestimado e são efetuadas previsões a cada iteração. A medida de erro total de previsão do modelo para *cross validation* é calculada como a média do erro para os k iterações. Quanto maior for o valor k, menor é a diferença entre os valores estimados e os valores reais, menor é o enviesamento. Assim, o objetivo é maximizar a precisão do modelo e minimizar a complexidade.

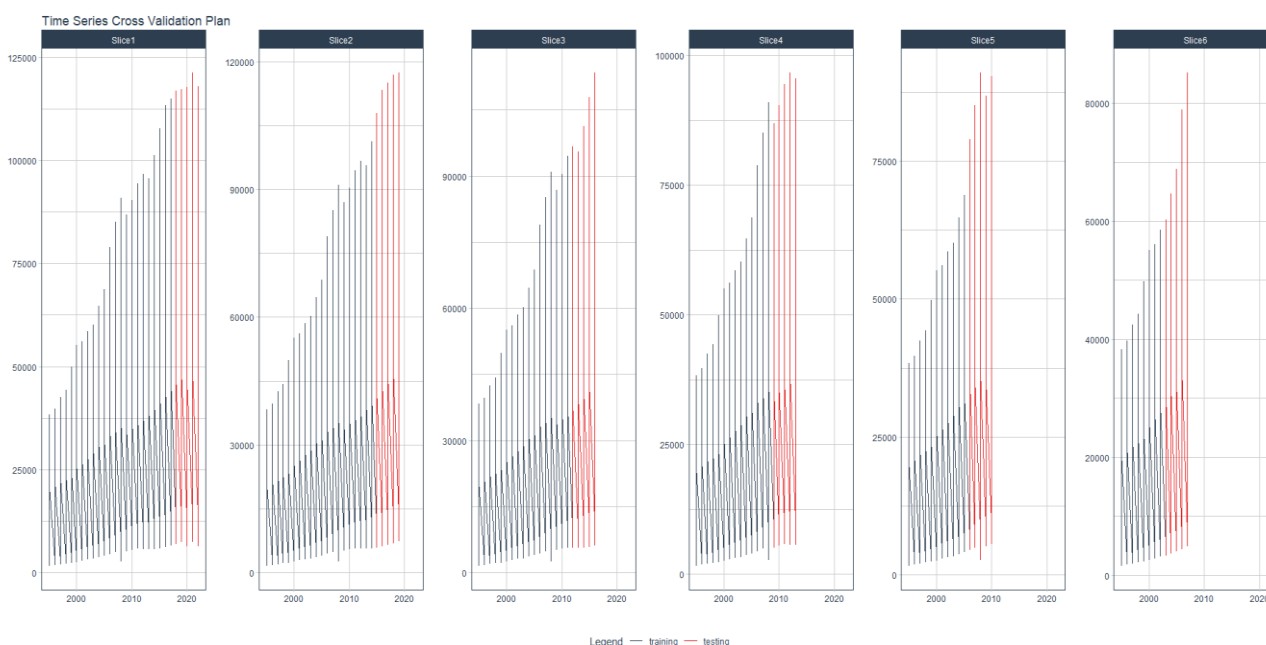


FIGURA 3.1: Representação gráfico do método de amostragem por validação cruzado de séries temporais.

3.4. Avaliação de precisão preditiva

A literatura acadêmica propõe várias medidas de desempenho preditivo, com base na amostra de teste. Em conformidade com Caiado (2016), destaco três medidas de avaliação dos erros de previsão. Seja y o PIBpc real, \hat{y} o PIBpc previsto e n o número de eventos por defeito na amostra. O erro quadrático médio (EQM) é calculado da seguinte forma

$$EQM = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 . \tag{3.43}$$

Quanto maior o valor de EQM, maior a diferença entre o valor real e o previsto e por isso menor precisão. Uma vez que EQM depende da média, considera-se que esta métrica é particularmente sensível a *outliers*. É possível reduzir a grandeza dos valores através da raiz do erro quadrático médio, $REQM = \sqrt{EQM}$.

O erro médio absoluto (EAM) é dado pela equação (3.36). Tanto o EQM como EAM são obtidos pela diferença média entre o PIBpc e estimadas, contudo o EAM pode refletir o tamanho do erro de forma mais intuitiva. A interpretação da medida do EAM, tal como a EQM, deve ter em consideração a escala implícita e por isso não é muito adequada para comparar dados de natureza e/ou periodicidade diferente.

O erro percentual absoluto médio (EPAM) permite obter uma percentagem média do erro preditivo face à real grandeza das observações. A medida de erro é definida da seguinte forma:

$$EPAM = \frac{1}{n} \sum_{t=1}^n \left| \frac{y_t - \hat{y}_t}{y_t} \right| \times 100. \quad (3.44)$$

Os modelos com valores inferiores a 100% para EPAM, proporcionam, em média, melhor previsão. A desvantagem desta métrica consiste em observações com zeros na amostra.

Também a correlação é um indicador importante, definido estatisticamente por

$$\rho_{y,\hat{y}} = \frac{Cov(y,\hat{y})}{\sqrt{Var(y)Var(\hat{y})}}. \quad (3.45)$$

Pela definição matemática, modelos com elevada correlação, apresentam baixos valores de variância e por isso maior precisão preditiva.

4.DADOS

4.1. Index of economic freedom database

O presente estudo consiste na criação de modelos preditivos para o produto interno bruto, em paridade de poder de compra, *per capita* (PIBpc). A escolha do PIBpc como variável de interesse consiste na necessidade de quantificar a atividade económica do país pelo seu custo de vida, por individuo nacional. Para tal, são utilizados dados em painel que contém 1 428 observações combinadas: 28 observações anuais, entre 1995 e 2022, para 51 países. A seleção dos países foi feita com base nos seguintes critérios: consistência de dados para o período em análise, inexistência de nulos e credibilidade da recolha dos dados. Os países utilizados no presente estudo encontram-se listados na Tabela A.4. em Apêndice.

Os dados em painel tem como principal fonte os dados disponibilizados pela *The Heritage Foudation* (2022), 19 indicadores disponibilizados no âmbito da publicação anual do índice de liberdade económica. Foram adicionados ainda 4 indicadores complementares, acrescentando

uma perspectiva mais social e ambiental da economia. A base de dados da OCDE, do Fundo Monetário Internacional, do Banco Mundial e da Conferência das Nações Unidas sobre Comércio e Desenvolvimento são outras fontes importantes para o enriquecimento da base de dados. Os indicadores que constam na modelação encontram-se na Tabela A.5.

Através da análise da Figura 4.1., é possível constatar forte tendência crescente do PIBpc dos vários países, porém podem ainda ser observados picos de decréscimo: em 2009, revelando o impacto da crise financeira de 2008; em 2020, devido à pandemia do COVID-19; em 2022, devido à guerra na Ucrânia. Importante realçar que a base de dados utilizada como referência diz respeito aos indicadores atuais e históricos para o cálculo anual do *Index of Economic Freedom*. Regra geral, o valor do PIBpc, associado ao índice do ano t , tem como principal fonte o *World Economic Outlook Database*, do mês de Abril do anterior, disponibilizado pelo Fundo Monetário Internacional.

Através da observação da Figura 4.1. Luxemburgo e Singapura destacam-se como os países com maior PIBpc anual. Porém, a maior taxa de crescimento entre 1995 e 2022 diz respeito à economia emergente da China, com um crescimento de 824% aproximadamente. É ainda possível constatar que em 1995, a economia com menor PIBpc era o Vietnã, e logo a seguir a Índia. Para 2022, são os Camarões o país com menor PIBpc, apesar da taxa de crescimento de 112% face a 1995.



FIGURA 4.1: Evolução do PIBpc para os 51 países.

4.1.1. Variáveis

A análise exploratória dos dados, é a etapa inicial do desenvolvimento de um modelo de *machine learning*, fundamental para o entendimento de cada variável e da relação da mesma com os as restantes variáveis. A Figura A.1 do Apêndice, revela a análise entre variáveis ao nível da correlação e da distribuição de frequências. Nos pontos seguintes serão analisadas as variáveis de forma individual. A população será analisada ao nível dos vários países. Para as restantes variáveis será utilizado um caso específico, Portugal, que apresenta valores em cada variável próximos da mediana no respetivo ano.

População

A Figura A.2. em Apêndice demonstra que a tendência da densidade populacional não é linear nos 51 países em análise. Apesar da tendência crescente na maioria dos países, é possível verificar que a Bulgária, Estónia, Geórgia, Hungria, Lituânia, Moldávia, Polónia, Roménia, Ucrânia, e ainda, Cabo Verde revela uma tendência decrescente. Este fenómeno deve-se ao forte movimento emigratório dos países do centro-leste europeu e dos países africanos. De destacar ainda a parábola côncava que configura a função de densidade populacional na Grécia, Portugal e Japão, devido à diminuição da natalidade.

Da interpretação obtida da expressão matemática do cálculo do PIBpc, é possível concluir que a variação PIBpc é bastante superior às variações da densidade populacional, em termos globais ao longo dos anos, resultando num maior acréscimo significativo do PIBpc.

Estado de direito

PropertyRights, *RuleofLaw* e *GovernmentIntegrity* são variáveis representativas do estado de direito (Figura A.3). Pela análise da Figura A.1. podemos constatar que existe uma alta correlação, acima dos 75%, entre estas variáveis. Em 2022, Portugal revelou o maior aumento dos últimos anos em *PropertyRights*. Apesar das fortes oscilações associadas a *RuleofLaw* e *GovernmentIntegrity* verificaram um acréscimo bastante significativo. Após a acentuada descida do valor do índice em 2020, o valor de *RuleofLaw* retomou aos valores normais. Relativamente a *GovernmentIntegrity* verificou-se dois picos negativos em 1997 e 2018, recuperando os valores nos anos seguintes.

Tamanho do governo

Esta seção tem em análise os gastos do governo como um todo, e ainda, setores, nomeadamente educação, saúde e meio-ambiente. Em Portugal, a variável *GovtExpenditure* revela uma tendência crescente de 1995 a 2010 e decrescente de 2011 a 2022 (Figura A.4). Apesar das oscilações, *GovtExpenditureEnvironmentalProtection* e *TotalHealthSpending* revelaram ter

o seu valor máximo em 2010, ano como já referi que também corresponde ao pico máximo da *GovtExpenditure*. No ano anterior à pandemia, 2019, os gastos públicos em saúde apresentaram o seu valor mínimo. Em oposição, no mesmo ano, 2019, a variável *GovernmentExpenditureEducation* revelou o seu valor máximo. De realçar que dos setores em análise, a educação é o que revela menor gasto público em comparação com os restantes, como é possível observar na figura seguinte.

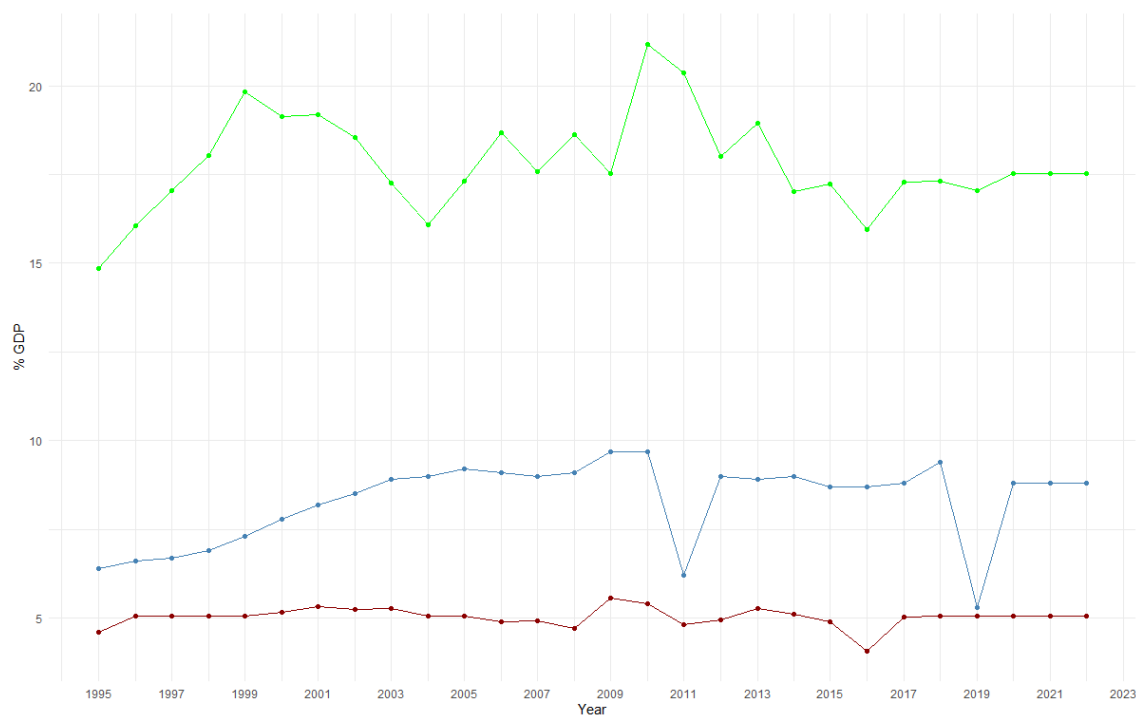


FIGURA 4.2: Evolução despesas públicas em percentagem do PIB, onde os pontos bordô, azul e verde correspondem respetivamente à educação, saúde e proteção ambiental.

Observando a Figura A.5, podemos observar o comportamento das variáveis *TaxBurden*, *FiscalHealth* e *PublicDebt* em Portugal. O índice da carga fiscal revela-se relativamente constante, porém verifica-se uma acentuada subida a partir de 2020. Relativamente ao índice de saúde fiscal, o pico mínimo foi verificado em 2017, verificando-se posteriormente uma acentuada subida, com valor máximo em 2021. Apesar da instabilidade do valor da dívida pública nos últimos 10 anos, a tendência revela-se crescente.

Eficiência regulamentar

A análise da eficiência regulamentar será subdividida em negócios e mercado de trabalho (Figura A.6.) e mercado monetário (Figura A.7.).

A evolução da variável *BusinessFreedom*, em Portugal, revela um crescimento entre 2005 e 2015 e posteriormente um decréscimo até ao ano atual. Em relação à variável *LaborFreedom*, verifica-se um repentino decréscimo a partir de 2008 até 2013, retomando em 2015 a valores

estáveis. A função de *Unemployment* apresenta o seu pico mínimo em 2000 e pico máximo em 2015. Os valores mais elevados de desemprego verificam-se entre 2012 e 2018, período de intervenção da troika em Portugal e de recuperação económica.

A função que descreve o comportamento de *MonetaryFreedom*, de Portugal revela fortes oscilações entre os 82 e os 87, apenas com exceção de 1995 e 1996 que revelam valores inferiores a 80. *Inflation* apresenta uma tendência decrescente com algumas flutuações, em particular 2008 devido à crise financeira.

Mercados abertos

As transações económicas de um país com o resto do mundo podem ser contabilizadas a partir do conceito de balança de pagamentos, nomeadamente em termos de balança corrente e balança financeira. O valor da própria balança corrente, a liberdade comercial e taxas aduaneiras são variáveis explicativas associadas ao primeiro componente da balança (Figura A.8.). Liberdade financeira, liberdade de investimento e FDI são variáveis explicativas associadas ao segundo componente anteriormente mencionado (Figura A.9.).

Em Portugal, *CurrentAccountBalance* tem sofrido algumas oscilações nos últimos anos, com uma forte queda entre 1995 e 2000, e ainda, entre 2020 e o presente ano. A variável *TariffRate* revela uma tendência decrescente até 2016 e uma subida posterior, que contrasta com *TradeFreedom*, uma vez que, revela valores crescentes com exceção dos 2 últimos anos.

Por sua vez, *FinancialFreedom* e *InvestmentFreedom* destacam-se por uma acentuada subida em 2009 e 1998, respetivamente, em relação ao ano anterior. Finalmente, o *ForeignDirectInvestment* (FDI) revela bastantes oscilações, apesar de manter-se bastante contante nos últimos anos. De destacar ainda que face ao PIB, o FDI apresenta valor máximo em 2012.

5. Resultados

5.1 XGBoost – Ajuste de hiperparâmetros

A Tabela 5.1 apresenta os hiperparâmetros obtidos pelo algoritmo de otimização de *grid search* com o objetivo de contribuírem para um melhor desempenho do modelo XGBoost.

Hiperparâmetros	Valor
Learning rate	0.2
Max tree depth	15
Gamma	0
Sub sample	0.7
Column sample	0.0008
Minimum child weight	1

TABELA 5.1: Hiperparâmetros utilizados para ajustar o modelo XGBoost.

5.2 Comparação da performance dos modelos

A tabela seguinte (Tabela 5.2) apresenta as medidas de precisão out-of-sample para os modelos de interpretação natural, *Time-fixed effects* (TFE), POLS (PO) e Árvore de decisão (DT), e ainda para o modelo de caixa preta, XGBoost (XGB). Foram ainda calculadas as variações percentuais obtidas das medidas de precisão entre dois modelos.

	TFE	PO	DT	XGB	PO-TFE(%)	DT-TFE(%)	XGB-TFE(%)
REQM	7567.83	16640.11	21178.79	12163.03	-55	-64	-38
EAM	5265.36	11437.81	15204.11	7691.47	-54	-65	-32
EPAM	22.31	40.45	37.79	18.69	-45	-41	19

TABELA 5.2: Comparação das medidas de precisão preditiva *out-of-sample* do PIBpc.

Podemos verificar que, relativamente a REQM e EAM, o modelo TFE é o modelo de previsão de dados de painel com maior precisão face à estimação do PIBpc com base em dados de teste. Porém, quando comparamos os modelos em termos de desempenho preditivo face ao EPAM, podemos verificar que XGBoost revela melhor desempenho do que TFE. Considerando a diversidade da natureza dos regressores e a ausência de zeros, EPAM constitui o melhor critério de avaliação dos erros de previsão.

5.3 Interpretabilidade Global

Esta seção pretende apresentar métodos de interpretação para os modelos de interpretabilidade natural e *black-box*, permitindo de alguma forma comparar os mesmos.

5.3.1 Modelos de interpretabilidade natural

Modelos lineares

A seleção do modelo mais adequado aos dados em painel, definidos por um atributo individual (*CountryName*) e temporal (*Year*), consiste na realização de testes estatísticos. Inicialmente, é feita a comparação entre o modelo de regressão dos dados agrupados (POLS) e o modelo de regressão de efeitos fixos (*within*) através do Teste F, no qual a hipótese nula é de que há igualdade nos interceptos e inclinações para todos os indivíduos, logo o modelo adequado é melhor ao modelo de dados agrupados. Uma vez que se verifica que o valor-p é inferior a 0,05, o modelo de efeitos fixos é melhor do que o modelo POLS. Tendo em conta o resultado, foi ainda aplicado o teste de Hausman que pretende comparar a adequação dos dados ao modelo de efeitos fixos e ao modelo de efeitos aleatórios, segundo o qual a hipótese nula é de os efeitos individuais c_i não ser correlacionado com os regressores x_{it} , isto é, se verificam-se efeitos

aleatórios. Como o valor-p é inferior a 0,05, o modelo de efeitos fixos é mais adequado comparativamente ao modelo de efeitos aleatórios. Foi ainda efetuado o teste F de efeitos individuais para verificar a existência de efeitos não observados de tempo ou individuais nos modelos de dados em painel. Uma vez que se verifica que o valor-p é inferior a 0,05, então verifica-se a existência de efeito tempo, *Time-fixed effects*. Apesar de não ser o modelo que melhor se adequa aos dados, o modelo POLS também foi explorado dada a maior simplicidade a aproximação aos restantes modelos em estudo.

Em termos de interpretabilidade, a influência de cada variável explicativa na previsão de cada modelo pode ser obtida através do método de estimação de efeitos parciais, tendo em conta a distribuição média de cada variável e cujo as restantes variáveis são mantidas constantes. Dado o contexto atual da economia internacional, a variável explicativa *Inflation* foi selecionada para análise mais profunda nos dois modelos. Na Figura A.10 em Apêndice podemos verificar que para o modelo POLS, a previsão tem uma variação negativa com o aumento da inflação. Assumindo um ponto específico como exemplo, se a inflação atingir os 2,3% então o valor obtido para a previsão é, em média, 24 680.181, *ceteris paribus*. Para o modelo *Time-fixed effects*, o impacto da variação da inflação revela uma tendência oposta à verificada pelo modelo POLS, isto é uma tendência positiva pouco significativa. Do mesmo modo, se a inflação atingir os 2,3%, o valor da previsão é, em média, 27 096.477 *ceteris paribus*.

A Figura A.11 apresenta o gráfico *classical feature importance* obtido pelo modelo POLS. As variáveis explicativas, *LaborFreedom*, *CurrentAccountBalance* e *GovernmentExpenditureEducation* apresentam maior nível de importância, 12.7, 12.1 e 7.55, respetivamente. Tal como *GovernmentExpenditureEducation*, também, *Unemployment* e *TariffRate* são os fatores que revelam um impacto negativo, com 5.50 e 5.11.

Segundo o método *permutation feature importance* para o modelo POLS (Figura A.12.a), *LaborFreedom*, *CurrentAccountBalance* e *PropertyRights* são as variáveis explicativas com maior importância tendo em conta a maior perda de erro quadrático médio, 11 134, 10 824 e 10 681, aproximadamente. Para o modelo *Time-fixed effects* (Figura A.12.b), *CountryName*, *Population* e *Year* são as variáveis com maior importância, com perdas aproximadas de 30 196, 18 756 e 12 018, respetivamente.

Árvore de decisão

O método mais simples de avaliação da importância das variáveis das variáveis explicativas num modelo de Árvore de decisão é através da observação da própria árvore, na qual é possível perceber que as variáveis no topo da árvore apresentam maior importância relativa. Para uma mesma amostra, a copa da árvore será sempre a mesma, porém para um modelo mais complexo,

o gráfico de árvore pode ser ajustado tal como é ilustrado pela Figura 5.1. A primeira questão prende-se com o facto de *RufeofLaw* ser inferior a 88.45. Se “sim”, é posteriormente perguntado se *TradeFreedom* é inferior a 80.75, se “não”, é perguntado se *TradeFreedom* é inferior a 85.95.

Se a segunda questão for se *TradeFreedom* é inferior a 80.75, e a resposta for “sim”, então é posteriormente questionado se *FiscalHealth* é igual ou superior a 69.3. Se “sim”, então o PIBpc esperado é de, aproximadamente, 8 975 dólares e o ramo termina; se “não”, então o PIBpc é 17 162 dólares aproximadamente. Se, por sua vez, *TradeFreedom* é igual ou superior a 80.75 então a terceira pergunta é se *RufeofLaw* é inferior a 60.96. Se “sim”, então o PIBpc esperado é aproximadamente 13 883 dólares, se “não”, então existe uma quarta pergunta. A quarta pergunta consiste em definir se *PublicDebt* é inferior a 37.475, se “sim”, então o PIBpc é 19 338 dólares aproximadamente, se “não”, então o PIBpc é 29 223 dólares aproximadamente.

Se a segunda questão for se *TradeFreedom* é inferior a 85.95 e a resposta for “sim”, então é posteriormente questionado se *CurrentAccountBalance* é inferior a 6.25. Se “sim”, a quarta questão é se *LaborFreedom* é inferior a 31.71. Se a resposta for “sim” então o PIBpc esperado é aproximadamente 26 483 dólares, se “não”, é aproximadamente 35 923 dólares. Se *LaborFreedom* é igual ou superior a 31.71 então o valor esperado do PIBpc corresponde a 44 647 dólares por aproximação. Se *TradeFreedom* é igual ou superior a 85.95, então a pergunta seguinte é se *TotalHealthSpending* é igual ou superior a 7.05, se a resposta for “sim” então a quarta pergunta é relativo a *TradeFreedom*, caso contrário, é relativo a *InvestmentFreedom*. Se *TradeFreedom* é inferior a 8.46 então o valor esperado do PIBpc é aproximadamente 45288 dólares, se for igual ou superior a 8.46, então o valor esperado é aproximadamente 63 040 dólares. Por sua vez, a quarta pergunta é se *InvestmentFreedom* é inferior a 92.5 e a resposta é “sim”, então o PIBpc é 77 268 dólares, por aproximação, se “não”, então o valor esperado do PIBpc é aproximadamente 103 504 dólares anuais, em média, por pessoa.

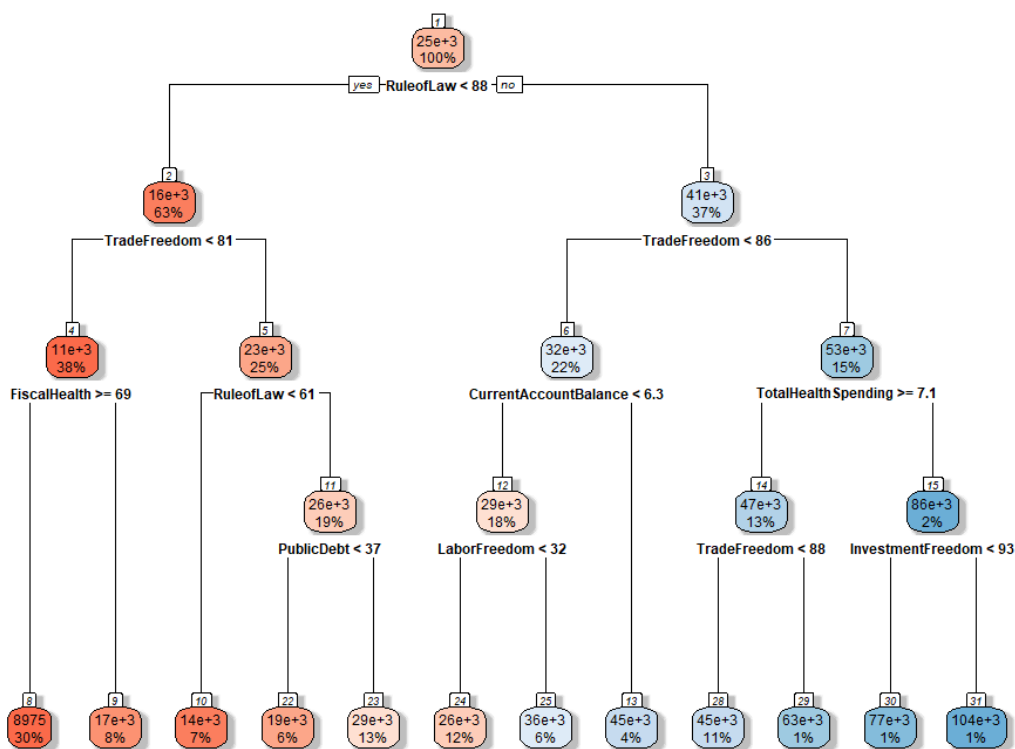


FIGURA 5.1: Diagrama de árvore.

Tendo em conta que a árvore completa é grande, é possível utilizar a redução da soma dos quadrados de cada divisão para analisar quais as variáveis mais importantes do modelo. A quantificação da importância revela vantagens, em particular, quando a mesma variável é utilizada em várias divisões da árvore. Os valores são padronizados em relação à variável explicativa mais importante, cujo valor é 100, e posteriormente são pontuados com base na redução da função de perda. A Figura A.11 em Apêndice mostra os resultados obtidos.

As variáveis *RuleofLaw*, *TotalHealthSpending*, *TradeFreedom*, *LaborFreedom* e *InvestmentFreedom* são identificadas como mais importantes em ambos os métodos. Existem ainda variáveis como *GovernmentIntegrity*, *PropertyRights*, *BusinessFreedom*, *MonetaryFreedom* e *TariffRate* que apesar de não constarem nas divisões iniciais da árvore, revelam importantes contribuições (Figura A.11) para a previsão sob o modelo de árvore.

5.3.2 Modelos black-box

XGBoost – eXtreme Gradient Boosting

Numa primeira fase, a compreensão global de XGBoost considera o método clássico de importância de cada variável. A Figura A.13 em Apêndice mostra os resultados obtidos para Frequência, Ganhos e Cobertura.

Através da análise dos gráficos, *Year* é a variável utilizada com maior frequência, em comparação a todas as outras variáveis. Contudo, em ganho e cobertura não revela uma importância significativa, encontra-se no nono e décimo terceiro lugar, respetivamente.

A variável *RuleofLaw* encontra-se no sétimo lugar das variáveis mais frequentes, revelando ser ainda a variável com maior ganho. Em termos de cobertura das observações, a variável encontra-se na quinta posição.

A variável que revela um maior impacto numa porção significativa das observações é *ForeignDirectInvestment*, isto é, com maior cobertura, e apesar não revelar um grande ganho, é o terceiro no ranking da importância face à frequência.

Nenhuma variável encontra-se simultaneamente no Top 3 de todas as métricas, revelando diferentes classificações entre as diferentes medidas e, conseqüentemente, não possibilitando uma comparação linear nem uma interpretação consensual sobre a avaliação das variáveis mais relevantes.

A Figura A.12.c em Apêndice, obtida da aplicação do método de *permutation feature importance*, apresenta as interações do modelo sobre a diminuição do desempenho do modelo. *GovernmentIntegrity* é a variável identificada como a mais importante, uma vez que leva a um aumento aproximado de 5 482 no REQM de previsão do modelo após a permutação dos valores do regressor. A segunda e a terceira variáveis mais importantes são *TotalHealthSpending* e a *RuleofLaw* com um aumento de 4 563 e 4 151, respetivamente, no REQM. Por oposição, as variáveis com menor importância são *CountryName* e *Year* com 0.57 aproximadamente, o que permite constatar que a permutação destas variáveis envolve uma mínima redução do erro.

O valor SHAP é um método alternativo que permite identificar a importância das variáveis baseando-se na magnitude das contribuições das regressões para a previsão. A variável com maior relevância é *GovernmentIntegrity*. Também, *PropertyRights*, *GovtExpenditure* e *RuleofLaw* revelam uma posição forte no *ranking* de importância, que vai de encontro com o obtido pelo método de *permutation feature importance*, com exceção apenas de *LaborFreedom*.

Para a análise dos efeitos marginais das variáveis mais importantes na previsão do modelo XGBoost, a Figura A.14. apresenta os gráficos de dependência parcial (PDP) para as quatro variáveis com maior valor médio SHAP, *GovernmentIntegrity*, *PropertyRights*, *GovtExpenditure* e *RuleofLaw*.

Através da análise PDP, observa-se que o valor da previsão do PIBpc é mais elevado nos cenários em que o índice para cada uma das quatro variáveis vai aumentando. Por outras palavras, a previsão é maior quando se verifica: (i) o aumento da integridade dos governos; (ii) o aumento

da eficácia dos direitos de propriedade; (iii) o controlo das despesas governamentais; (iv) o aumento da regulamentação. É possível constatar a tendência crescente do índice de cada variável com o valor da previsão.

5.3.3 Interpretabilidade global entre modelos

A Tabela 5.3 apresenta uma síntese das cinco variáveis mais importantes para cada modelo. A análise do quadro permite-nos observar que *GovernmentIntegrity*, *LaborFreedom*, *PropertyRights*, *RuleofLaw* E *TotalHealthSpending* são variáveis consensuais em quase todos os modelos, uma vez que são definidas como importantes para pelo menos três modelos.

	PO- PFI	TFE- PFI	TR- CFI	XGB- PFI	XGB- SHAP
BusinessFreedom			5°		
CountryName		1°			
CurrentAccountBalance	2°				
FiscalHealth					
GovtExpenditure					3°
GovernmenttExpenditureEducation	5°				
GovernmentIntegrity	4°		2°	1°	1°
LaborFreedom	1°	4°		3°	
Population		2°			
PropertyRights	3°		3°	5°	2°
RuleofLaw			1°	4°	4°
TariffRate		5°			
TotalHealthSpending			4°	2°	5°
Year		3°			

TABELA 5.3: Comparação das variáveis selecionadas mais importantes nos modelos.

De realçar que para XGBoost, *GovernmentIntegrity* é a variável mais importante para ambos os métodos de interpretabilidade e corresponde à única variável que corresponde simultaneamente ao Top 5 de ambos os métodos. Relativamente ao método de valor SHAP, as três variáveis com maior importância para o modelo *black-box* também são bastante relevantes para os modelos de interpretabilidade natural.

Os modelos de interpretabilidade natural permitem uma compreensão mais rápida e mensurável do impacto de cada variável explicativa sobre os modelos, pelo que a interpretação global não exige métricas adicionais fornecidas pelo modelo.

Em oposição, os modelos *black-box* requerem maior esforço para o cálculo de métricas mais profundas e complexas. Para além disso, verificam uma maior diversidade de métodos e métricas de avaliação de importância, o que só por si pode também constituir um fator de incerteza sobre o método mais adequado. Uma clara definição do objetivo da análise pode facilitar a escolha do método a adotar, por exemplo analisar a contribuição de cada variável na previsão ou medir o impacto de cada variável na redução do erro. Apesar do objetivo parecer semelhante, cada método concentra-se na análise de um aspeto específico, pelo que *SHAP value* e *permutation feature importance* são, respetivamente, os métodos mais adequados para análise da previsão e da perda do erro.

	PO- PDP	TFE- PDP	XGB- PDP
GovernmentIntegrity	(+)	(-)	(+)
PropertyRights	(+)	(+)	(+)
GovtExpenditure	(-)	(+)	(+)
RuleofLaw	(+)	(+)	(+)
TotalHealthSpending	(+)	(-)	(+)

TABELA 5.4: Sinal dos efeitos das variáveis mais importantes identificadas pelo SHAP.

Relativamente ao sinal dos efeitos obtido pelo método de PDP para as cinco variáveis mais importantes segundo a metodologia SHAP (Tabela 5.4), conclui-se que, apenas os sinais para *PropertyRights* e *RuleofLaw* são iguais entre os três modelos. Para PO e XGB, os sinais relativos a *GovernmentIntegrity* e *TotalHealthSpending* são idênticos. A variável *GovtExpenditure* apresenta maior discordância entre os modelos, e isso deve-se ao facto de as despesas públicas serem, com peso e medida, uma variável positiva e negativa face ao desenvolvimento económico de um país. O resultado revela alguma concordância com a teoria económica, uma vez que, o aumento da confiança nas entidades governamentais, a eficácia dos direitos de propriedade como apoio ao investimento privado, a melhoria da regulamentação, e, o maior controlo dos gastos públicos e apoio ao setor da saúde, são medidas que apresentam um impacto positivo na previsão do PIBpc.

CONCLUSÃO

O principal objetivo deste trabalho consiste em comparar a relação entre qualidade de precisão e interpretação de modelos interpretáveis (modelo POLS, *Time-fixed effects* e Árvore de decisão) e um modelo *black-box* (*Extreme Gradient Boosting*). Este tópico desperta grande interesse para organizações financeiras e políticas, nacionais e internacionais, uma vez que a previsão do PIB tem um grande impacto nas decisões e políticas a implementar.

Por meio da utilização da base de dados *Index of Economic Freedom*, como principal fonte, o presente trabalho demonstra que o modelo XGBoost apresenta uma melhor precisão preditiva, tendo em conta o erro percentual absoluto médio. Apesar do modelo *Time-fixed effects* ter um menor erro face às restantes medidas de avaliação, o erro percentual absoluto médio é a métrica que melhor se adequa aos dados, uma vez que, traduz a percentagem média do erro em relação à grandeza efetiva das observações. Também, o ajustamento adequado dos hiperparâmetros constitui uma vantagem do modelo XGBoost face aos restantes.

O presente trabalho, no que diz respeito à interpretabilidade global dos modelos, procurou compreender o funcionamento interno de cada um dos modelos selecionados. Foram implementados vários métodos de interpretabilidade dos quais destaco, *Partial Dependence Plot* para POLS *model* e *Time-fixed effects model*, e a análise do diagrama para Árvore de decisão. Para XGBoost foram aplicados *Permutation Feature Importance*, *SHAP values* e ainda *Partial Dependence Plot*.

A extração das regras de funcionamento dos modelos de interpretabilidade natural é mais simples, não exigindo um grande esforço adicional. Por outro lado, os modelos *black-box* revelam maior precisão, porém exigem maior esforço na compreensão de métricas não observáveis e bastante menos intuitivas. A aplicação de métodos de interpretabilidade garante maior transparência e confiança através da desconstrução de modelos mais complexos. Deste modo, obtemos estimativas de maior precisão e elevada interpretabilidade sobre os fatores inerentes à variável macroeconómica, proporcionando uma clara vantagem comparativa e maior eficácia das políticas adotadas.

REFERÊNCIAS BIBLIOGRÁFICAS

- Box, G.E.P., Jenkins, G.M. e Reinsel, G.C. (1994), *Time Series Analysis, Forecasting and Control*, (3), Prentice Hall, Englewood Cliffs, 47-125.
- Caiado, J. (2016), *Métodos de Previsão em Gestão - Com Aplicações em Excel*, (2), Edições Sílabo, Lda., Lisboa, 34-57 e 204-254.
- Carriero, A., Clark, T. E. e Marcellino, M. (2015), 'Bayesian VARs: Specification choices and forecast accuracy', *Journal of Applied Econometrics*, 30 (1), 46–73.
- Chen, T. e Guestrin, C. (2016), 'Xgboost: A scalable tree boosting system', 785–794.
- Ciccarelli, M., Ortega, E. e Valderrama, M.T., (2016), 'Commonalities and cross-country spillovers in macroeconomic-financial linkages', *The B.E. Journal of Macroeconomics*, 16 (1), 231–275.
- Coulombe, P. G., Leroux, M., Stevanovic, D. S. e Surprenant, S. (2020), 'How is Machine Learning Useful for Macroeconomic Forecasting?'. University of Pennsylvania, Pennsylvania, e Université du Québec à Montréal, Montréal.
- Craven, M. e Shavlik, J. W. (1996), 'Extracting tree-structured representations of trained networks', *Advances in Neural Information Processing Systems*, 8, MIT Press, Cambridge, 24–30.
- Diebold, F.X. (1997), 'The Past, Present and Future of Macroeconomic Forecasting', *Journal of Economic Perspectives*, 12, 175-192.
- Diebold, F.X. (1998), 'Elements of Forecasting in Business, Economics, Government and Finance', South-Western College Publishing, Cincinnati.
- Fair, R.C. (1984), 'Specification, Estimation, and Analysis of Macroeconometric Models', Harvard University Press, Cambridge.
- Fisher, A., Rubin, C. and Dominici, F. (2018), 'Model class reliance: Variable importance measures for any machine learning model class, from the 'Rashomon' perspective'.
- Friedman, J. H. (2001), 'Greedy function approximation: A gradient boosting machine.', *The Annals of Statistics*, 29 (5), 1189–1232.
- Geweke, J. (1977), 'The Dynamic Factor Analysis of Economic Time-Series Models', North-Holland, Amsterdam, 365-383.
- Granger, C.W.J. (1969), 'Investigating Causal Relations by Econometric Models and CrossSpectral Methods', *Econometrica*, 37, 424-438.
- Greenwell, B., Boehmke, B. e McCarthy, A. (2001), 'A simple and effective model-based variable importance measure', arXiv preprint arXiv:1805.04755.

International Monetary Fund (IMF). Obtido em março 2021 e maio 2022, de IMF Datamapper Datasets: <https://www.imf.org/external/datamapper/datasets>.

Joseph, A. (2019), ‘Parametric Inference with Universal Function Approximators’.

Kearns, M. (1988), ‘Thoughts on hypothesis boosting (tech. rep.)’, University of Pennsylvania, Pennsylvania.

Kuan, C. M. e White, H. (1994), ‘Artificial neural networks: An econometric perspective. *Econometric Reviews*’, 13 (1).

Lee, T. H., White, H. e Granger, C. W. J. (1993), ‘Testing for neglected nonlinearity in time series models: A comparison of neural network methods and alternative tests’, *Journal of Econometrics*, 56 (3), 269–290.

Lucas, R.E. (1972), ‘Expectations and the Neutrality of Money’, *Journal of Economic Theory*, 4, 103-124.

Lucas, R.E. (1976), ‘Econometric Policy Evaluation: A Critique’. *Carnegie-Rochester Conference Series on Public Policy*, 1(1): 19-46. Amsterdam:North-Holland,

Lundberg, S. e Lee, S. (2017), ‘A unified approach to interpreting model predictions’, *Advances in neural information processing systems*.

Matos, S. M. (2020), ‘Interpretable Models of Loss Given Default’, (Dissertação de Mestrado). ISEG, Universidade de Lisboa.

Medeiros, M. C., Teräsvirta, T. e Rech, G. (2006), ‘Building Neural Network Models for Time Series: A Statistical Approach’, *Journal of Forecasting*.

Miller, Tim. (2017), ‘Explanation in artificial intelligence: Insights from the social sciences’, *Artificial Intelligence*, 267, 1-43.

Molnar, C. (2022), *Interpretable Machine Learning*, github. Disponível em: <https://christophm.github.io/interpretable-ml-book/>

Ribeiro, M., Singh, S. e Guestrin, C. (2016), ‘“Why should I trust you?”: Explaining the predictions of any classifier’, 97–101.

Sargent, T.J. e Sims, C. (1977), ‘Business Cycle Modeling Without Pretending to Have Too Much a Priori Theory’, *New Methods of Business Cycle Research*, Federal Reserve Bank of Minneapolis, Minneapolis.

Schapire, R. E. (1990), ‘The strength of weak learnability’, 5, 197–227.

Shapley, L. (1953), ‘Contributions to the theory of games’, *Contributions to the Theory of Games II*, Princeton University Press, Princetonpp, 307–317.

- Sigrist, F. (2018), ‘Gradient and newton boosting for classification and regression’.
- Simionescu M., Dobeš K., Brezina I. e Gaal A. (2016), ‘GDP rate in the European Union: simulations based on panel data models’, *Journal of International Studies*, 9 (3), 191-202.
- Sims, C.A. (1972), ‘Money, Income and Causality’, *American Economic Review*, 62, 540- 552.
- Sims, C.A. (1980), ‘Macroeconomics and Reality’, *Econometrica*, 48, 1-48.
- Slutsky, E. (1927), ‘The Summation of Random Causes as the Source of Cyclic Processes’, *Econometrica*, 5, 105-146.
- Stock, J. H. e Watson, M. W. (1999), ‘A Comparison of Linear and Nonlinear Univariate Models for Forecasting Macroeconomic Time Series’, *Cointegration, Causality, and Forecasting: A Festschrift in Honour of Clive W.J.*, Oxford University Press, Oxford, 1–44.
- Swanson, N. R. e White, H. (1997), ‘A Model Selection Approach to Real-Time Macroeconomic Forecasting Using Linear Models and Artificial Neural Networks’, *The Review of Economics and Statistics*, 79 (4), 540–550.
- Organisation for Economic Co-operation and Development (OECD). Obtido em abril 2021 e maio 2022, de OECD.Stat: <https://stats.oecd.org/>.
- Tashman, L. J. (2000), ‘Out-of-sample tests of forecasting accuracy: An analysis and review’, *International Journal of Forecasting*, 16 (4), 437–50.
- Taylor, J. (1993), ‘Macroeconomic Policy in a World Economy: From Econometric Design to Practical Operation’, Norton, New York.
- The Heritage Foundation, *2022 Index of Economic Freedom*. Obtido em março 2021 e junho 2022, de Index of Economic Freedom: <https://www.heritage.org/index/>.
- The World Bank. Obtido em abril 2021 e maio de 2022 em Data: <https://data.worldbank.org/indicator/>.
- Trapletti, A., Leisch, F. e Hornik, K. (2000), ‘Stationary and Integrated Autoregressive Neural Network Processes’, *Neural Computation*, 12 (10), 2427–2450.
- UNESCO. Obtido em março 2021, de UIS.Stat: <http://data.uis.unesco.org/>.
- Wooldridge, J. M. (2002), *Econometric Analysis of Cross Section and Panel Data*, 1, The MIT Press, Cambridge, 170-172 e 247–291.
- World Health Organization (WHO). Obtido em março 2021, de Global Health Expenditure Database: <https://apps.who.int/nha/database/Select/Indicators/en>.
- Yang, C., Rangarajan, A. e Ranka, S. (2018), ‘Global model interpretation via recursive partitioning’.

Yule, G.U. (1927), 'On a Method of Investigating Periodicities in Disturbed Series, with Special Reference to Wolfer's Sunspot Numbers', *Philosophical Transactions A*, 226.

Zhao, Q. e Hastie, T. (2019), 'Causal interpretations of black-box models', *Journal of Business & Economic Statistics*, 272-281.

APÊNDICE

Hyperparameter	Definition	Proposed setting
Learning rate	Shrinks the weights predicted by each tree. This is a fixed value in order to avoid overfitting and thus make the model more conservative.	0.01, 0.05, 0.1, and 0.2
Minimum child weigh	Gives the minimum number of instances required in a child node. If this value increases, it makes the model more conservative.	1,2,4 and 6
Maximum tree dep	Indicates the maximum amount of edges between the root node of a regression tree and its nodes. Increasing this value will make the model more complex and more likely to overfit.	4 to 14 with step size 2
Column sample percentage	Subsample ratio of columns when constructing each tree. Sub-sampling will occur once in every boosting iteration.	0.5, 0.9 and 1
Sub sample percentage	sample ratio of the training instances. Setting it to 0.5 means that XGBoost would randomly sample half of the training data prior to growing trees and this will prevent overfitting. Sub-sampling will occur once in every boosting iteration.	0.5, 0.9 and 1
Gamma	States the minimum loss reduction needed in order for a tree to make a new split. If the value for gamma is higher, trees become more shallow.	0 to 8, with step size 2

TABELA A.1: Hiperparâmetros XGBoost.

Input:
Trained model f
feature matrix X
target vector y
error measure $L(y, f(X))$
Process:
Estimate the original model error $e_{orig} = L(y, f(X))$ (e.g. mean squared error)
For each feature $j = 1, \dots, p$ do:
Generate feature matrix X_{perm} by permuting feature j in the data X
Estimate error $e_{perm} = L(y, f(X_{perm}))$ based on the predictions of the permuted data
Output:
Calculate permutation feature importance $PFI_j = e_{perm} / e_{orig}$. Alternatively, the difference can be used: $PFI_j = e_{perm} - e_{orig}$.
Sort features by descending PFI_j

TABELA A.2: Pseudocódigo do método *permutation feature importance* baseado em Fisher, Rudin, and Dominici (2018).

Input:
the unique predictor values $x_{11}, x_{12}, \dots, x_{1k}$
Process:
for $i \in 1, 2, \dots, k$
copy the training data and replace the original values of x_1 with the constant x_{1i}
compute the vector of predicted values from the modified copy of the training data
compute the average prediction to obtain $\bar{f}_1(x_{1i})$
Output:
the estimated partial dependence values $\bar{f}_1(x_{11}), \bar{f}_1(x_{12}), \dots, \bar{f}_1(x_{1k})$

TABELA A.3: Pseudocódigo baseado em Greenwell et al. (2001) para construção da dependência parcial da característica x_1 .

Lista de países em análise:

Albânia	France	Paraguay
Argélia	Georgia	Peru
Argentina	Germany	Philippines
Arménia	Greece	Poland
Austrália	Hungary	Portugal
Áustria	Iceland	Romania
Bélgica	India	Russia
Bolívia	Indonesia	Saudi Arabia
Botswana	Iran	Singapore
Brasil	Ireland	Slovakia
Bulgária	Israel	Slovenia
Camarões	Italy	South Africa
Canada	Jamaica	Spain
Cabo Verde	Japan	Sweden
Chile	Lithuania	Switzerland
China	Luxembourg	Thailand
Colômbia	Mexico	Tunisia
Costa Rica	Moldova	Turkey
Croatia	Mongolia	Ukraine
Czech Republic	Morocco	United Kingdom
Denmark	Netherlands	United States
Ecuador	New Zealand	Vietnam
Estonia	Nigeria	
Finland	Norway	

TABELA A.4: Lista de países presentes nos dados de painel para a variável CountryName.

Lista de indicadores macroeconómicos:

Year - ano

CountryName – país

Population – população (milhões)

Estado de direito:

PropertyRights – índice de direitos de propriedade

RuleofLaw – índice de eficiência jurídica

GovernmentIntegrity - índice de integridade governamental

Tamanho do governo:

GovtExpenditure – índice de despesa governamental

GovernmentExpenditureEducation – despesa pública em educação (% PIB)

TotalHealthSpending – despesa pública em saúde (% PIB)

GovtExpenditureEnvironmentalProtection – despesa pública em proteção ambiental (% PIB)

TaxBurden – índice de carga fiscal

FiscalHealth – índice de saúde fiscal

PublicDebt - dívida publica (% PIB)

Eficiência regulamentar:

BusinessFreedom - índice de liberdade empresarial

LaborFreedom – índice de liberdade laboral

Unemployment - taxa de desemprego

MonetaryFreedom – índice de liberdade monetária

Inflation - taxa de inflação

Abertura de mercado:

TradeFreedom – índice de liberdade comercial

CurrentAccountBalance - balança corrente

TariffRate - Taxa aduaneira

InvestmentFreedom – índice de liberdade investimento

FDI - investimento direto estrangeiro (% PIB)

FinancialFreedom – índice de liberdade financeira

TABELA A.5: Lista de variáveis explicativas incluídas nos dados de painel em estudo.

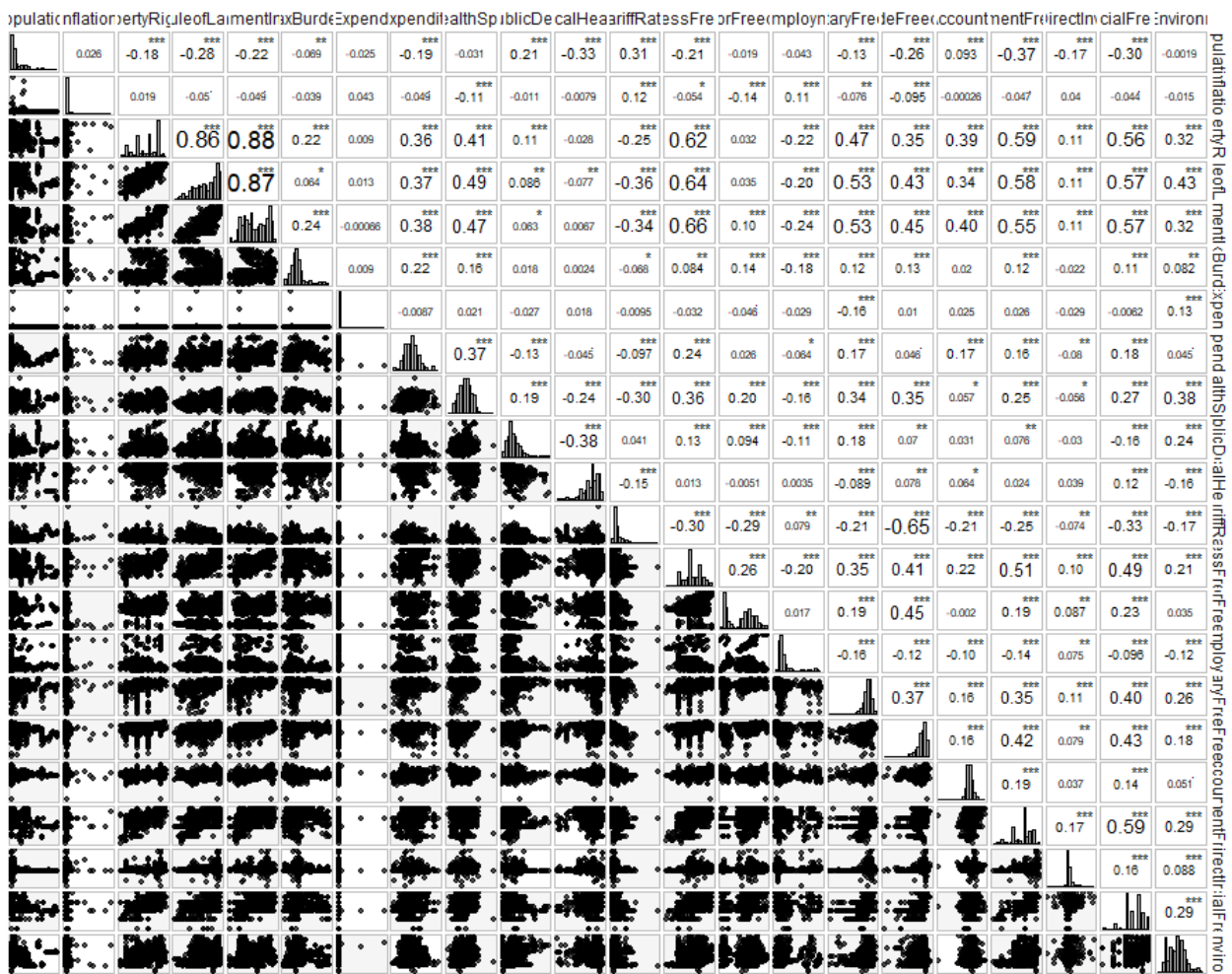


FIGURA A.1: Correlação das variáveis explicativas.

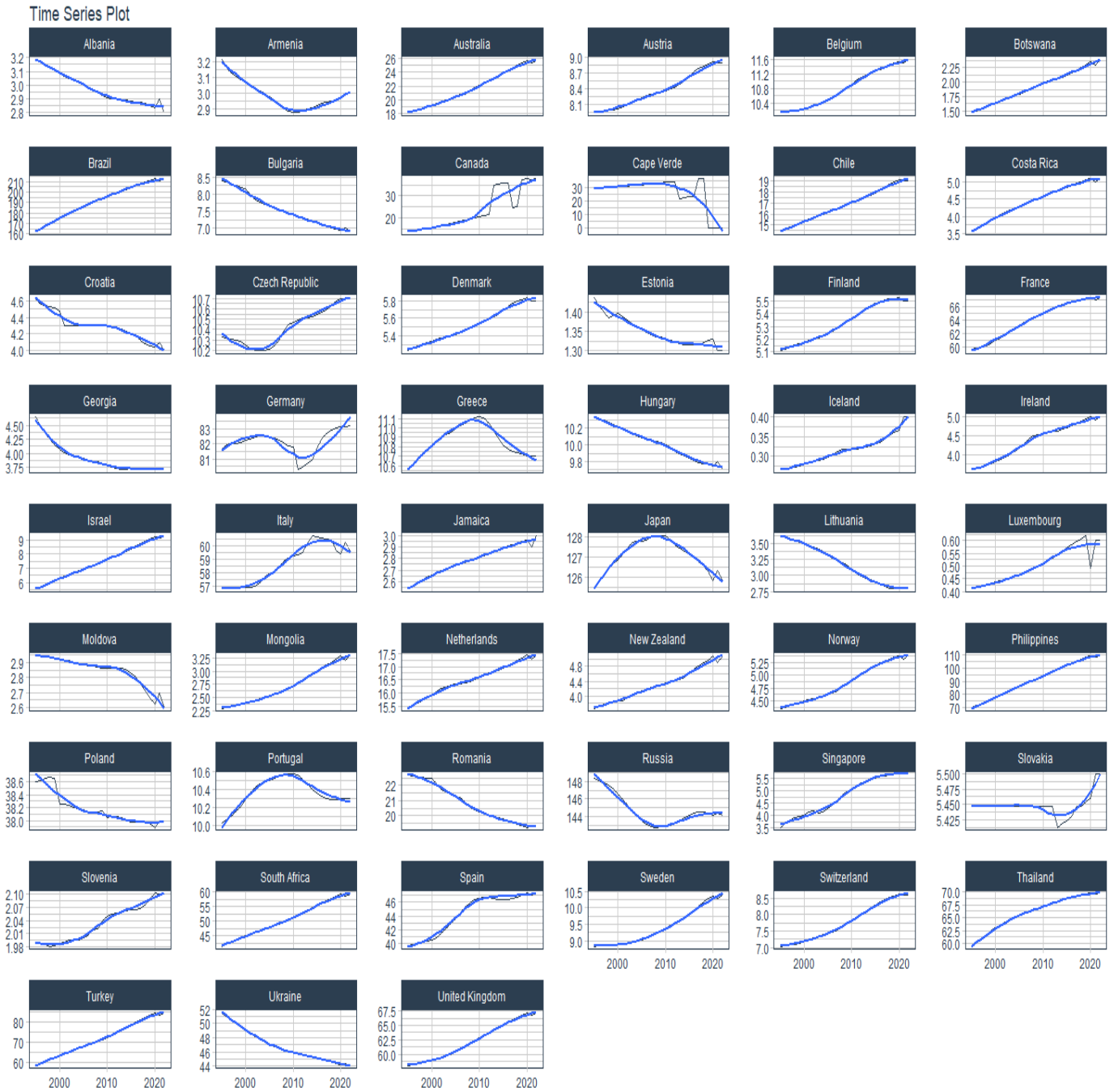


FIGURA A.2: Evolução da população nos diferentes países.

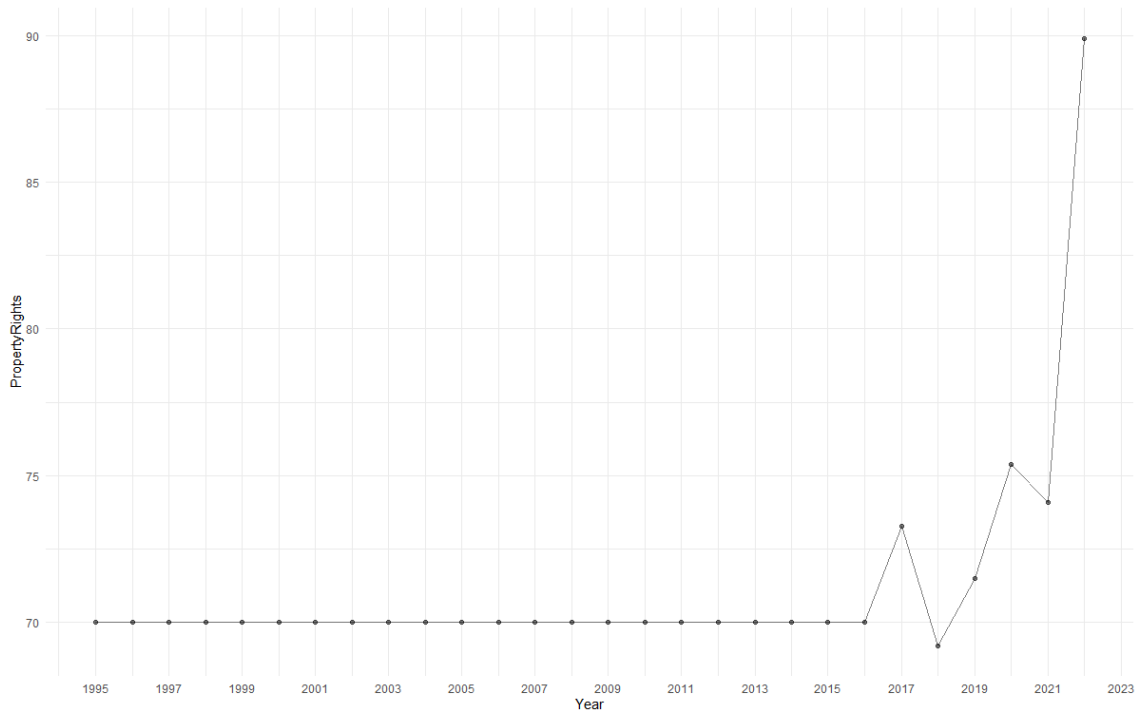


FIGURA A.3.a: Evolução do índice de direitos de propriedade de Portugal.

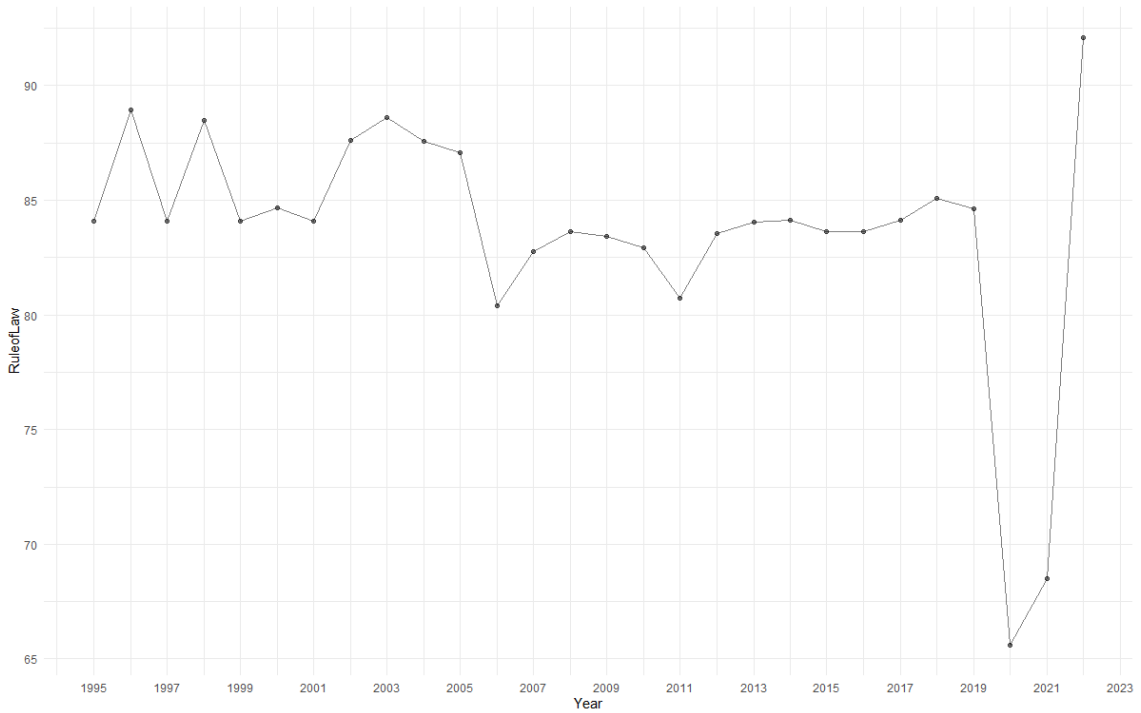


FIGURA A.3.b: Evolução do índice de regulamentação de Portugal.

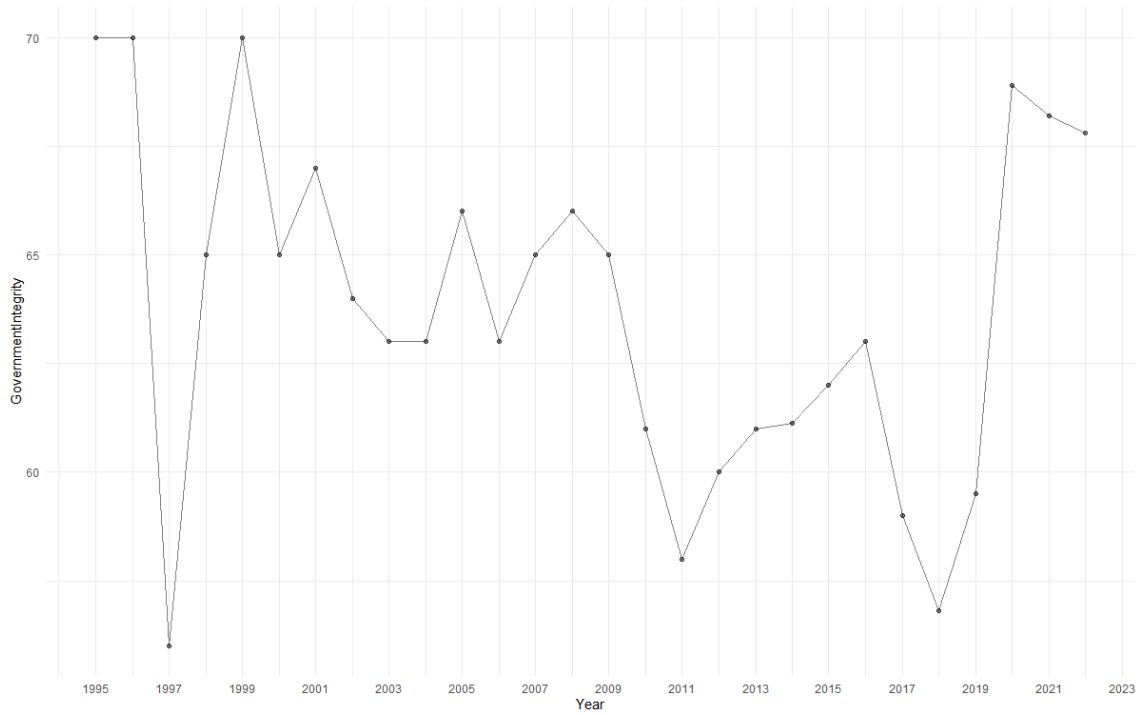


FIGURA A.3.c: Evolução do índice de integridade governamental de Portugal.

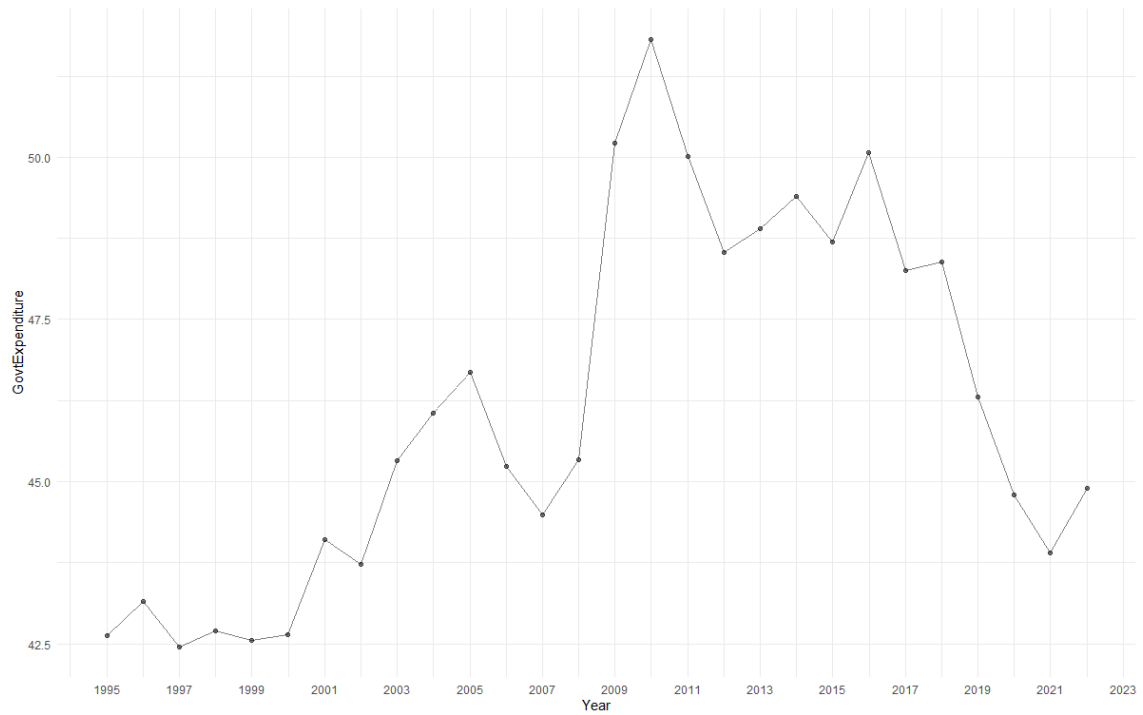


FIGURA A.4: Evolução do índice de despesas governamentais de Portugal.

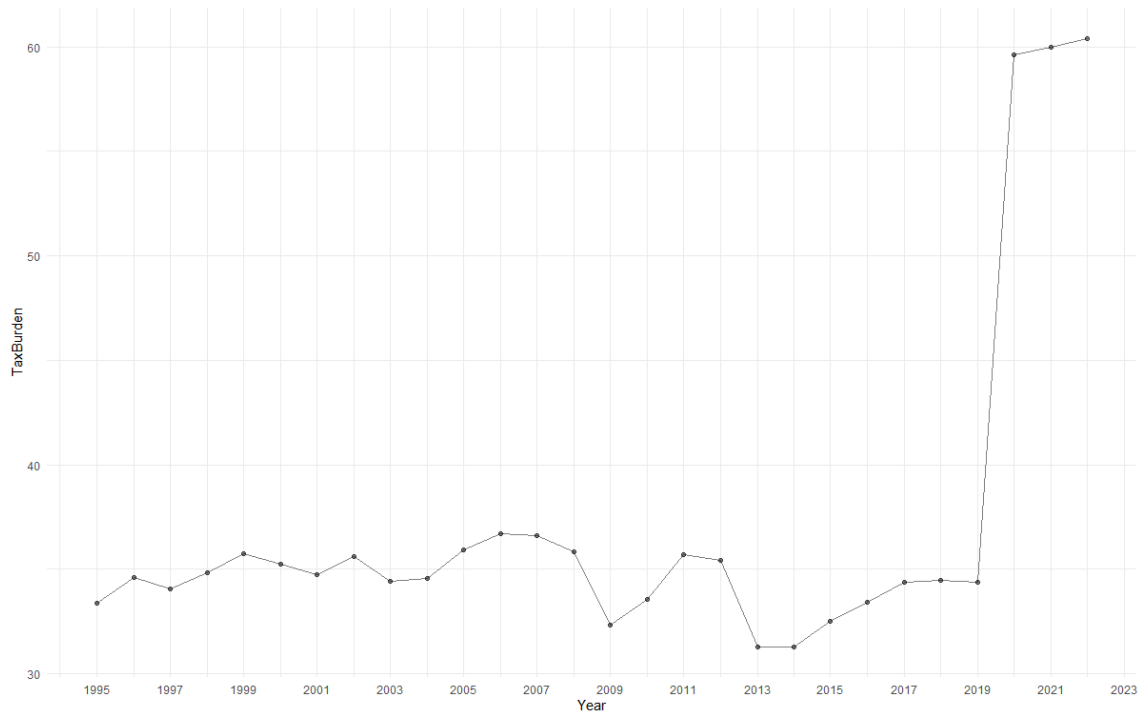


FIGURA A.5.a: Evolução do índice de carga fiscal de Portugal.

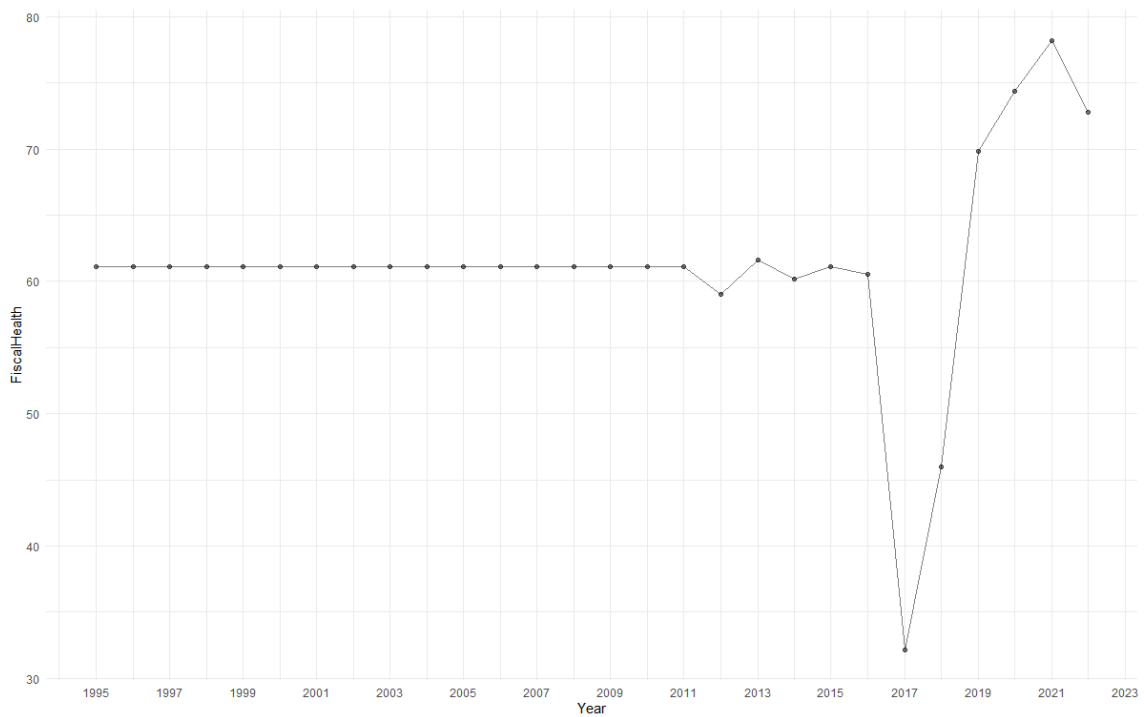


FIGURA A.5.b: Evolução do índice de saúde fiscal de Portugal.

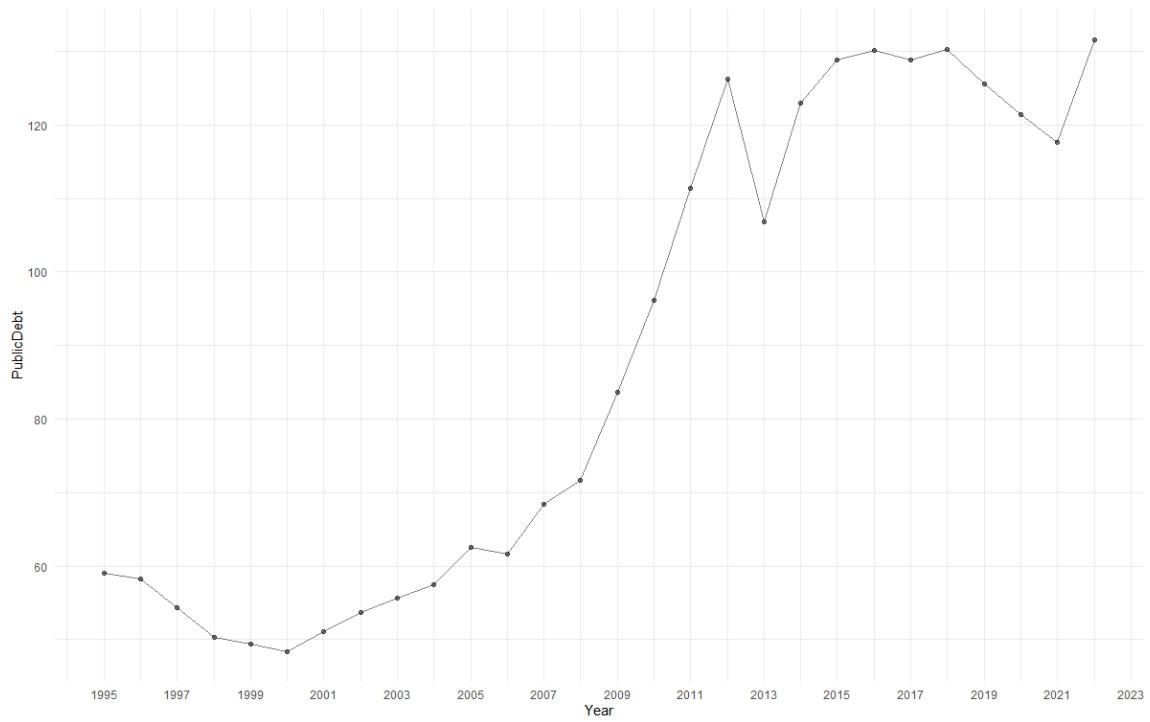


FIGURA A.5.c: Evolução da dívida pública (% PIB) em Portugal.

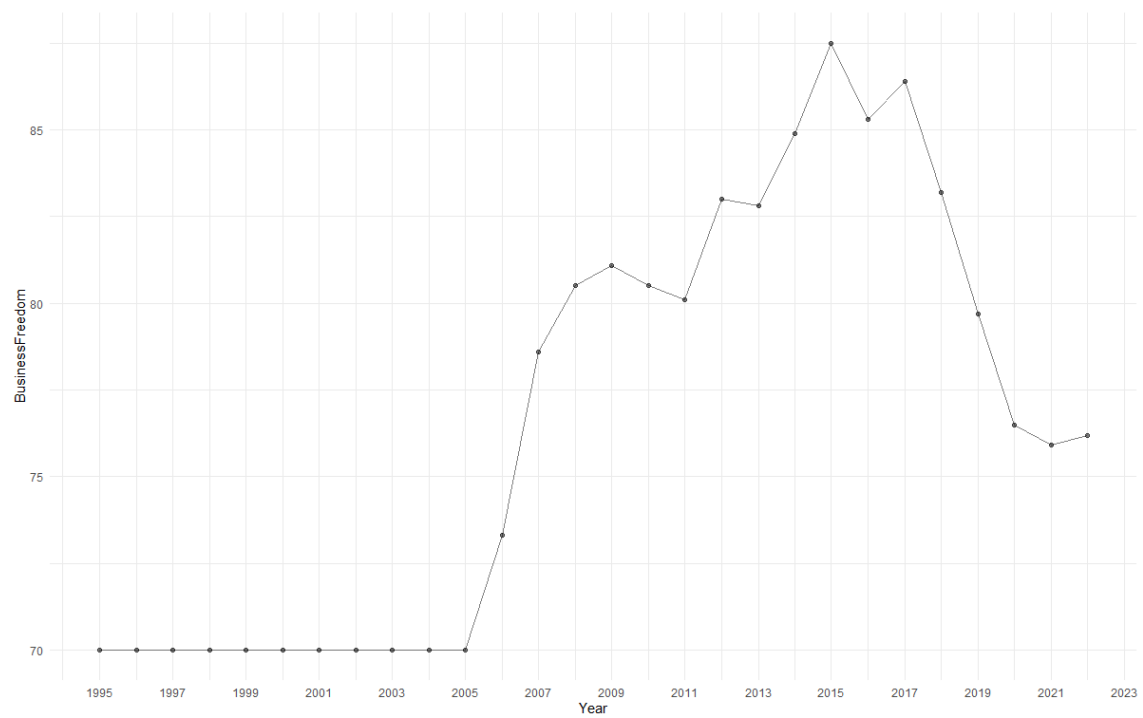


FIGURA A.6.a: Evolução do índice de liberdade de negócio de Portugal.

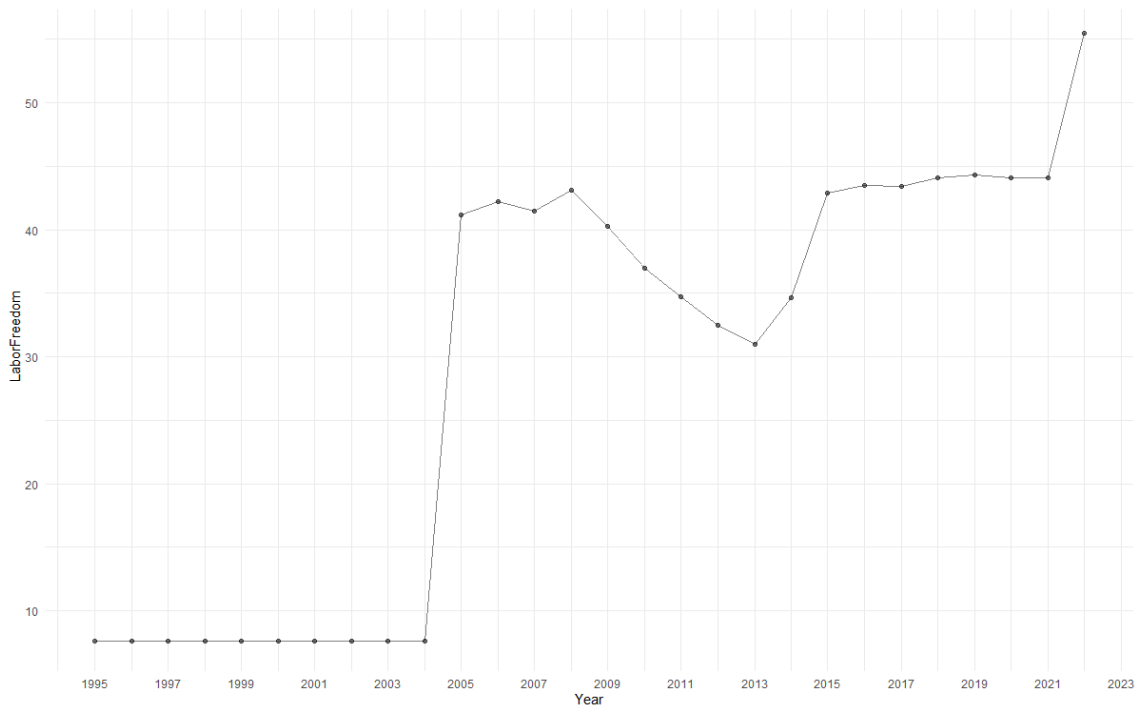


FIGURA A.6.b: Evolução do índice de liberdade laboral de Portugal.

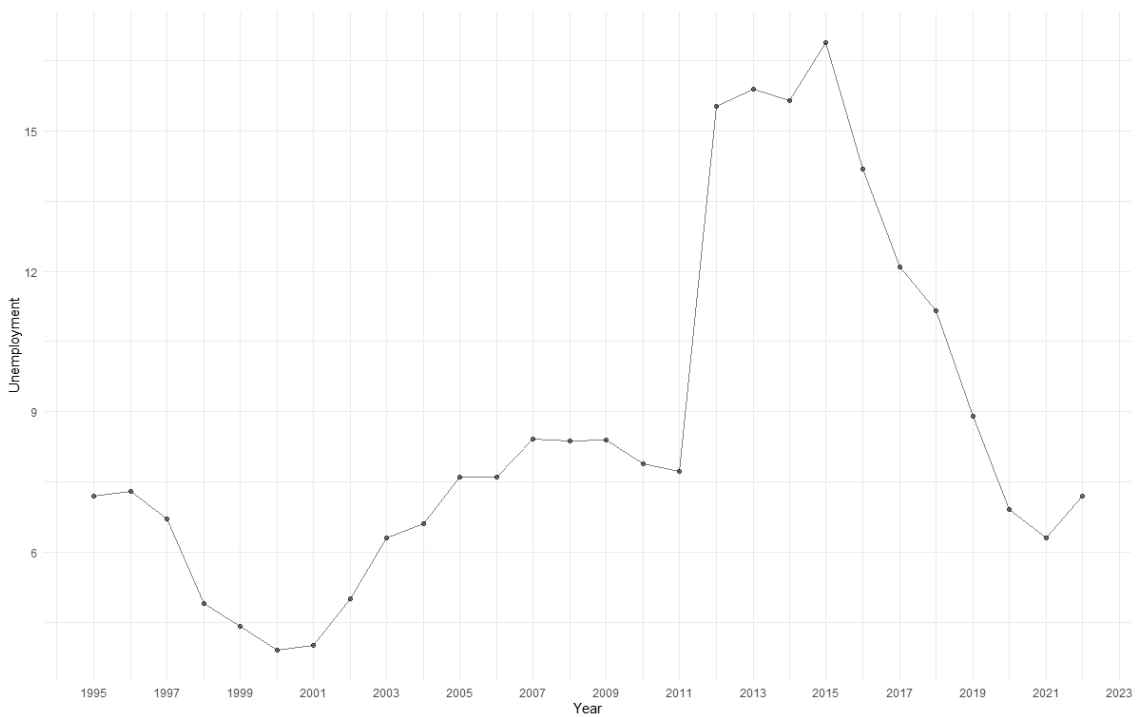


FIGURA A.6.c: Evolução da taxa de desemprego em Portugal.

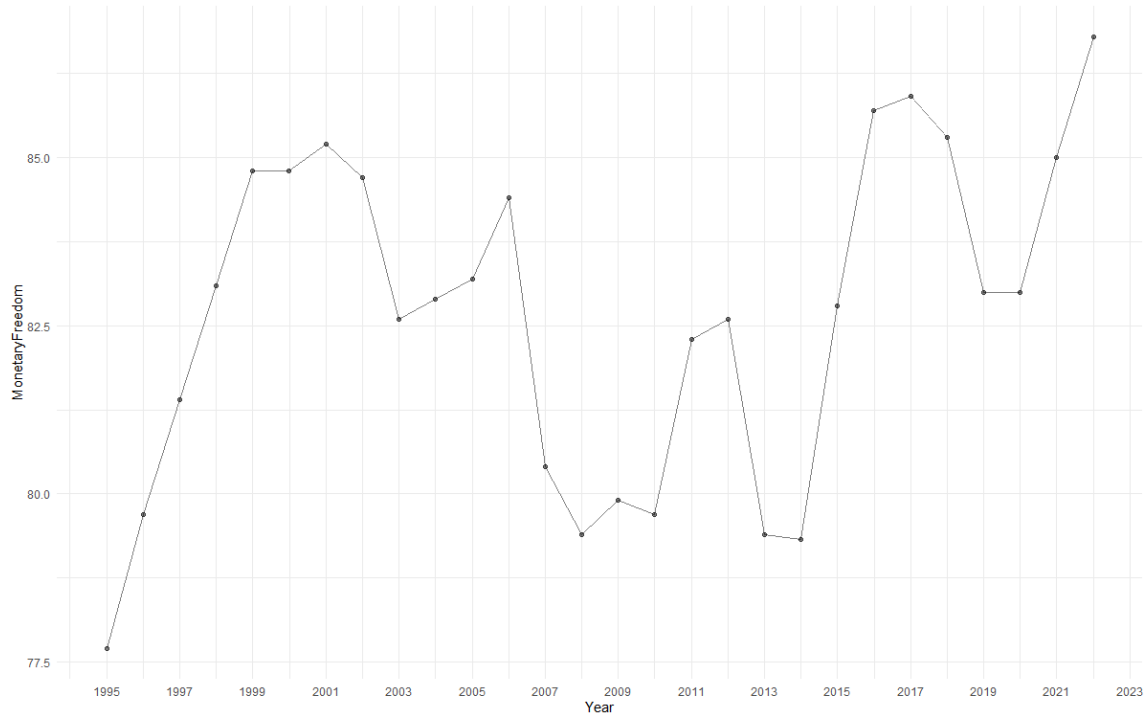


FIGURA A.7.a: Evolução do índice de liberdade monetária de Portugal.

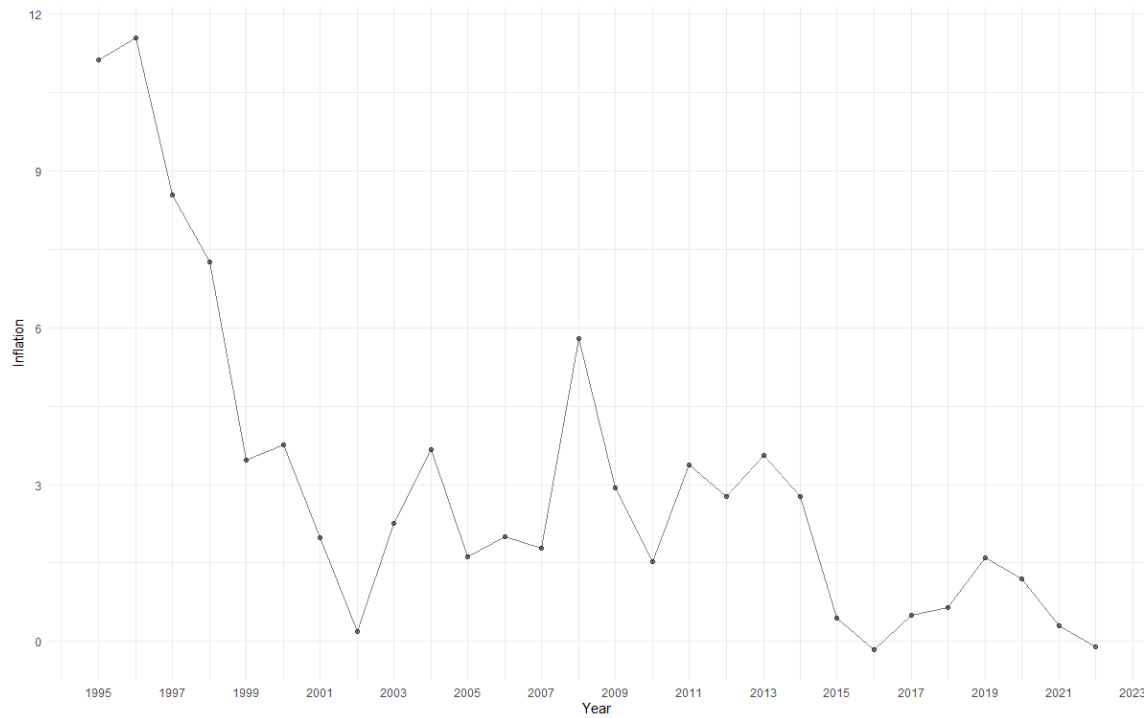


FIGURA A.7.b: Evolução da taxa de inflação em Portugal.

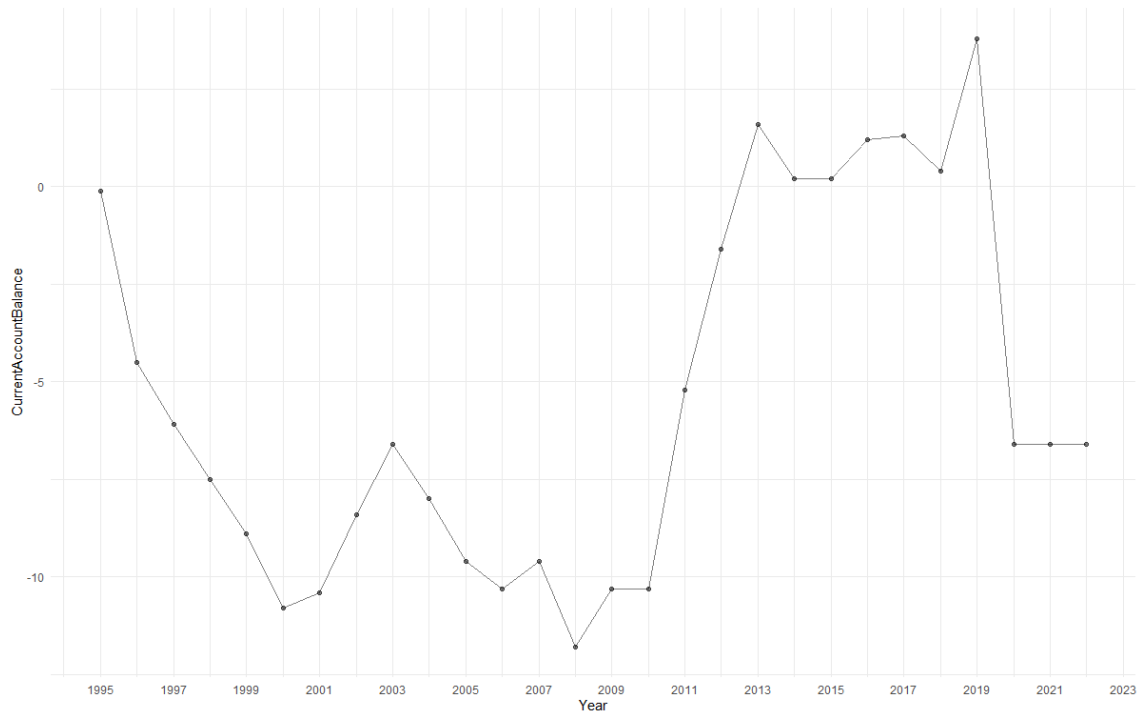


FIGURA A.8.a: Evolução da balança conta corrente de Portugal.

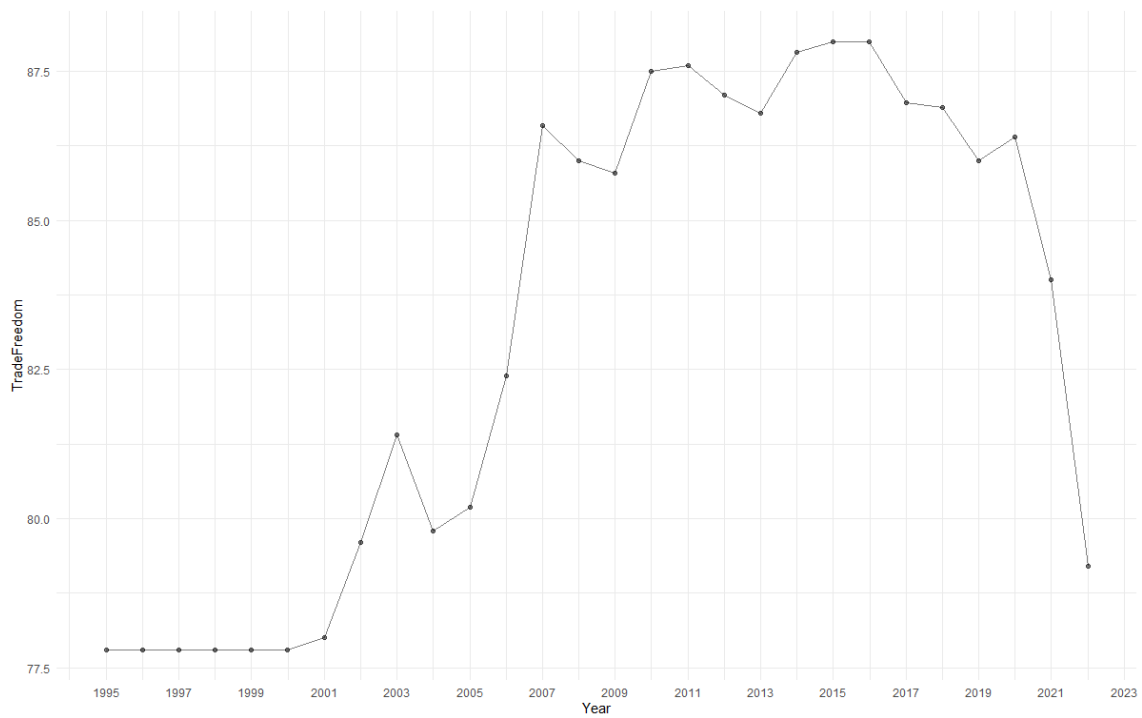


FIGURA A.8.b: Evolução do índice de liberdade comercial de Portugal.

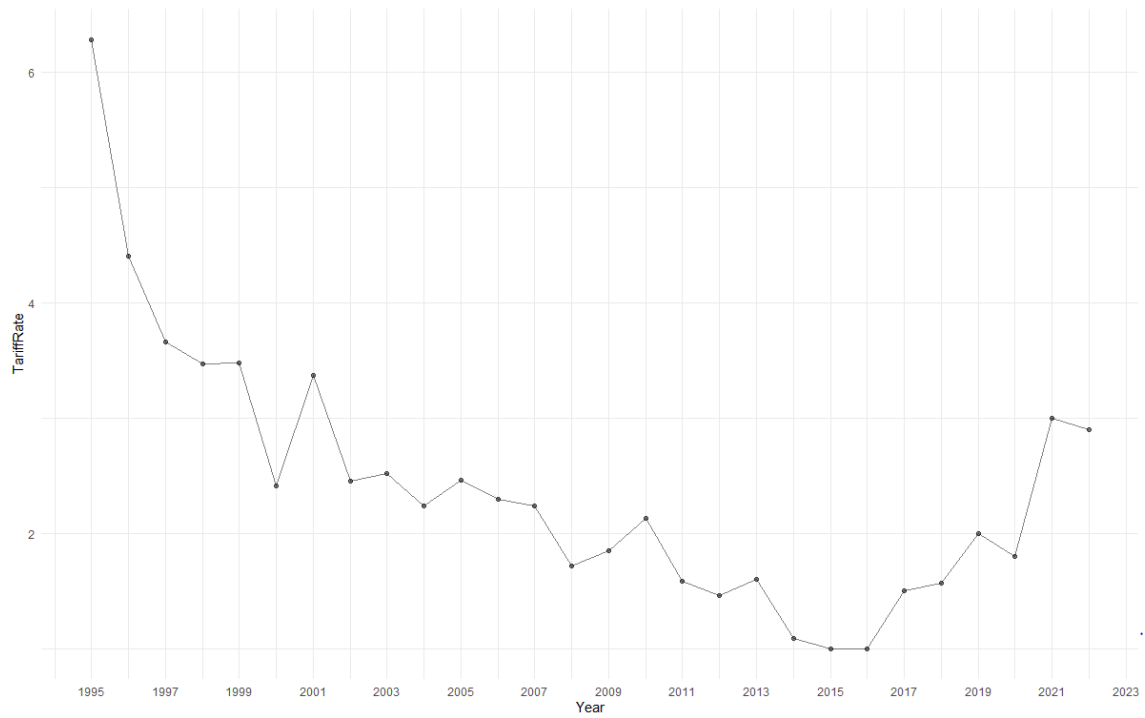


FIGURA A.8.c: Evolução do valor da taxa alfandegária de Portugal.

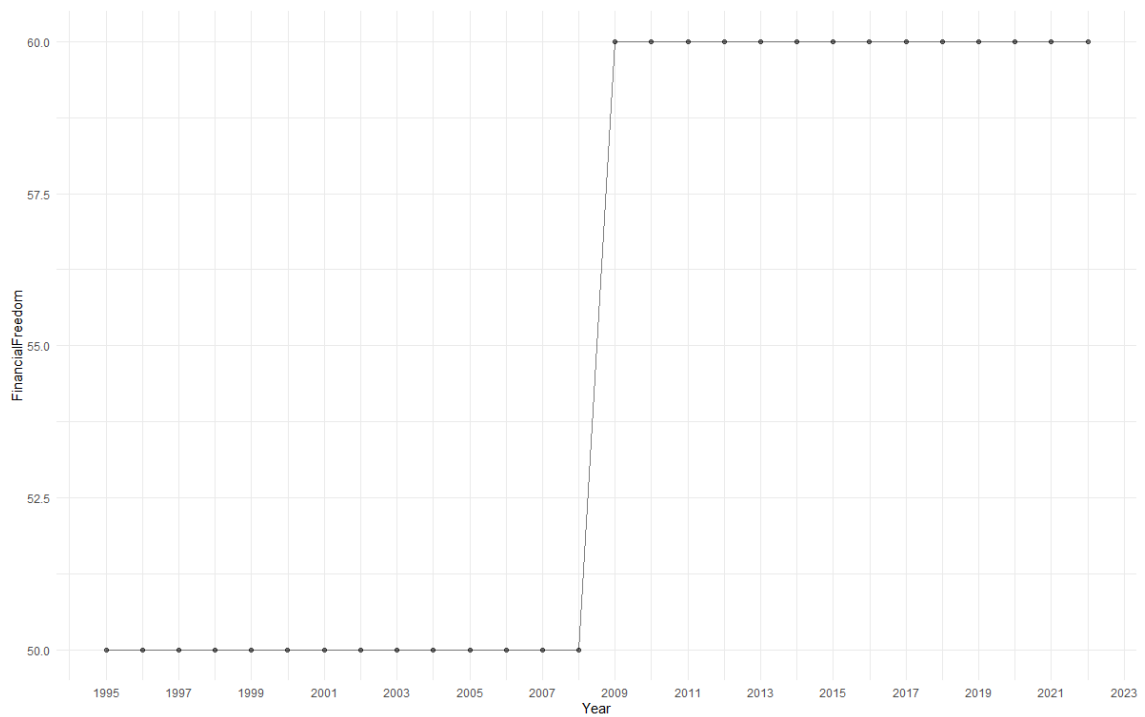


FIGURA A.9.a: Evolução do índice de liberdade financeira de Portugal.

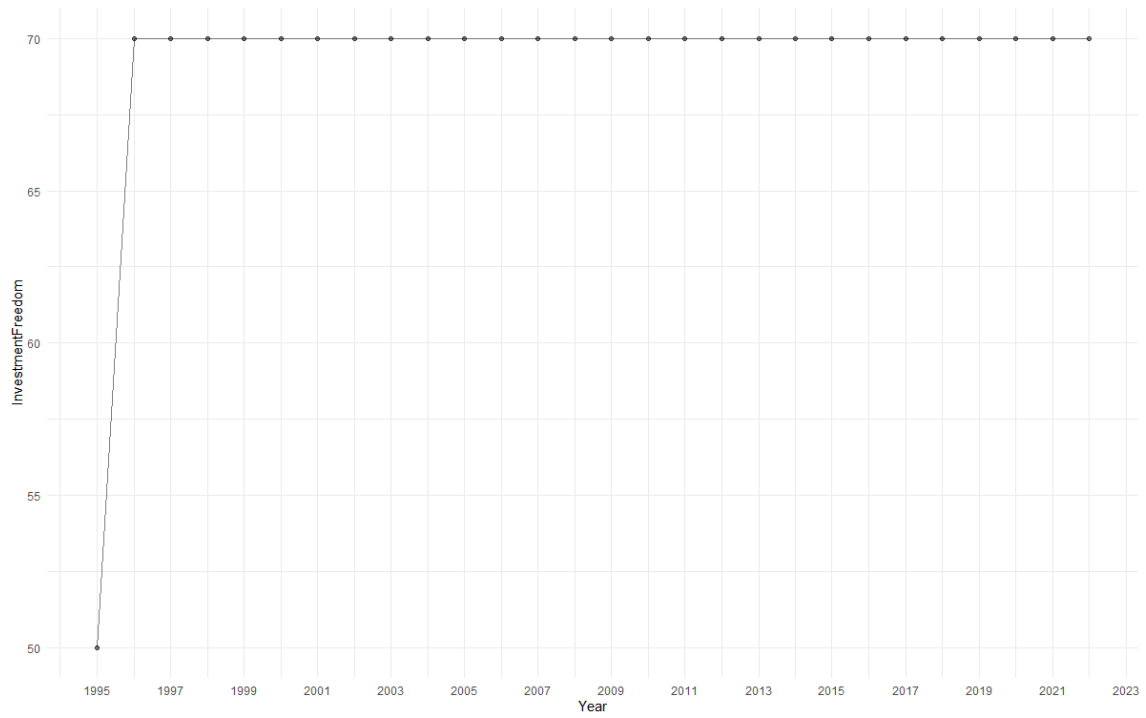


FIGURA A.9.b: Evolução do índice de liberdade de investimento de Portugal.

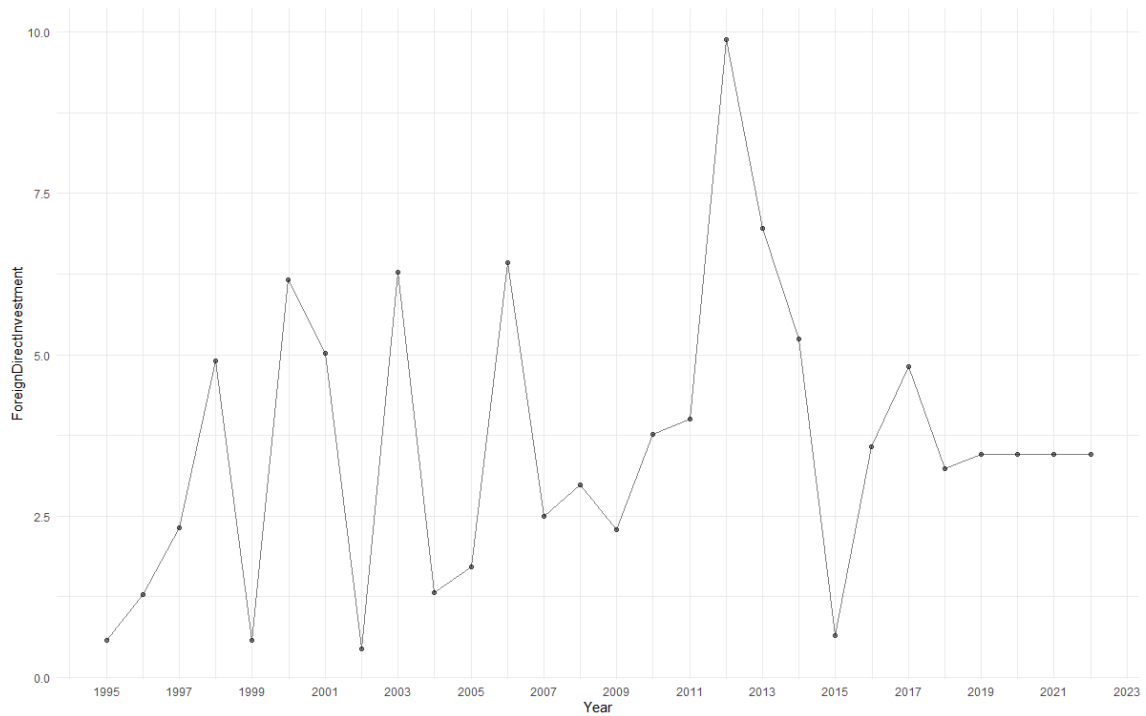


FIGURA A.9.c: Evolução do investimento direto estrangeiro (% PIB) em Portugal.

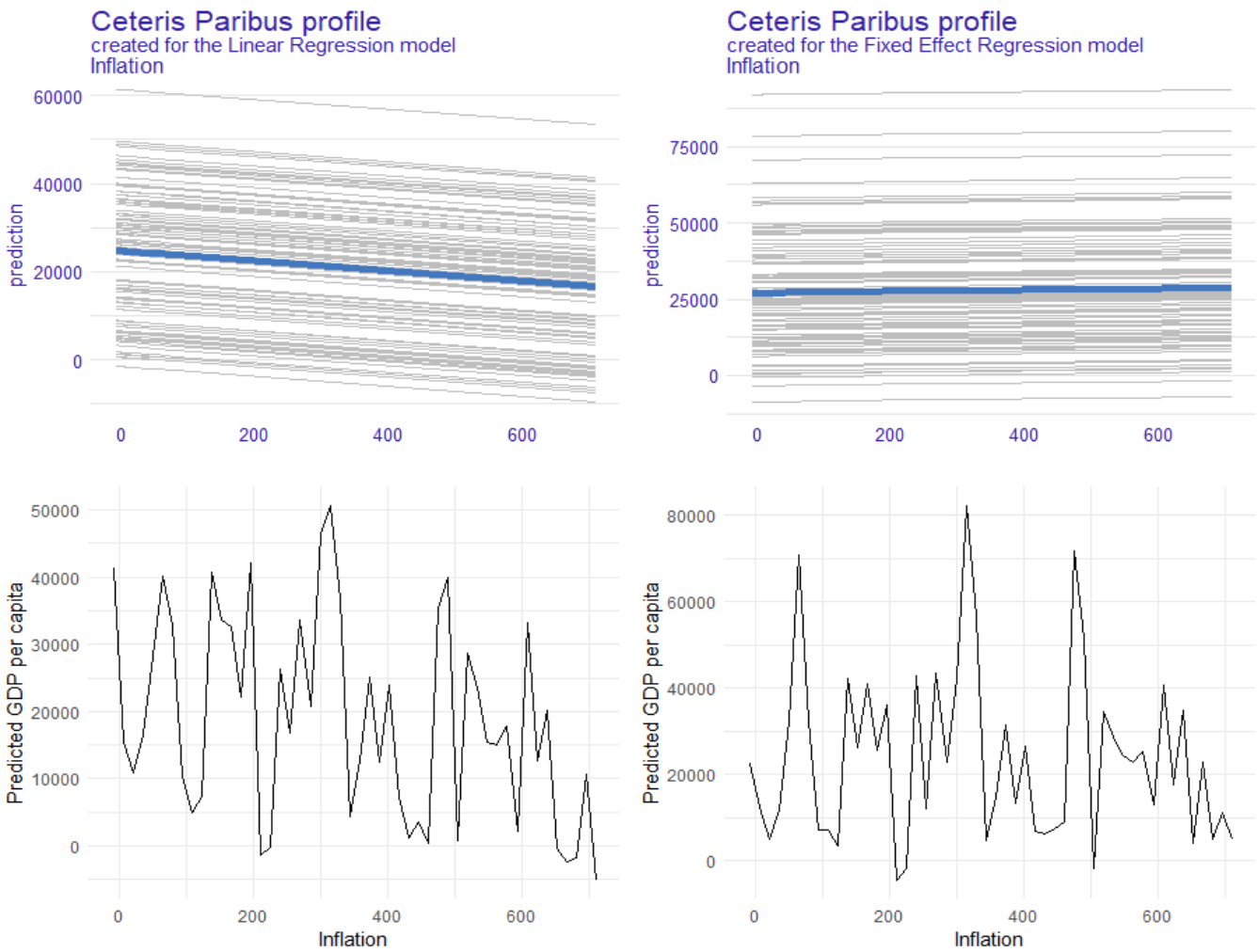


FIGURA A.10: Método de PDP aplicado à inflação.

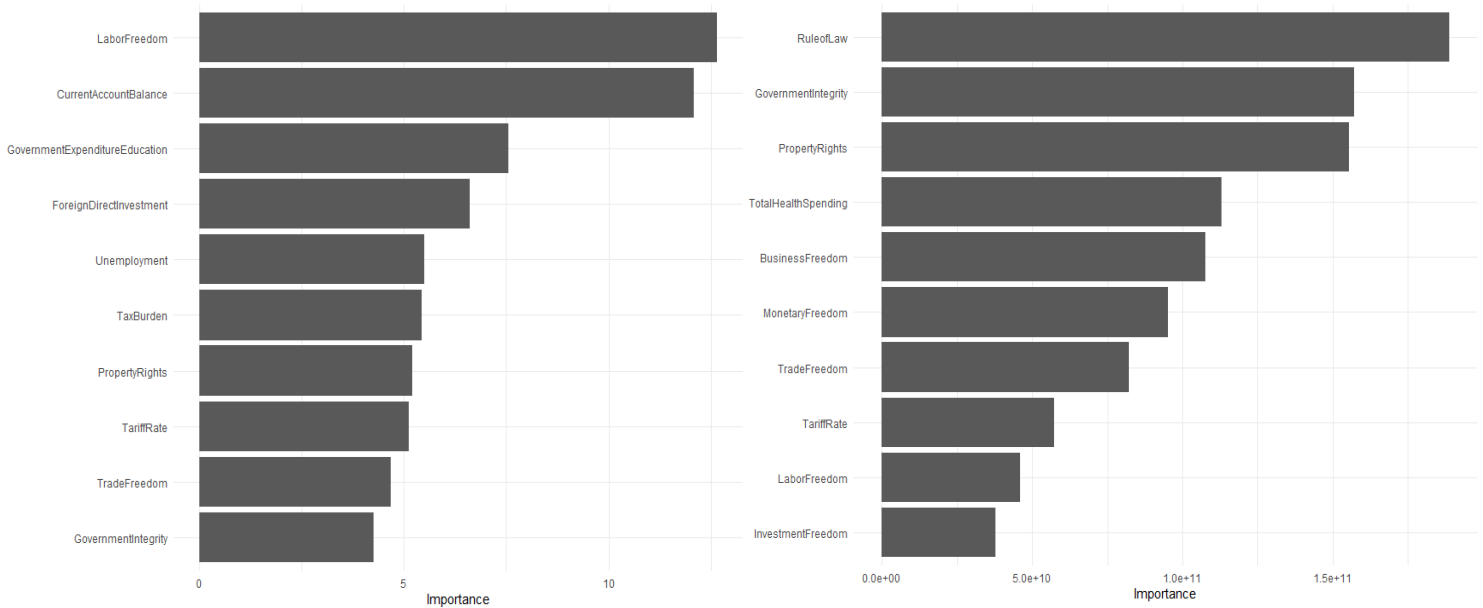
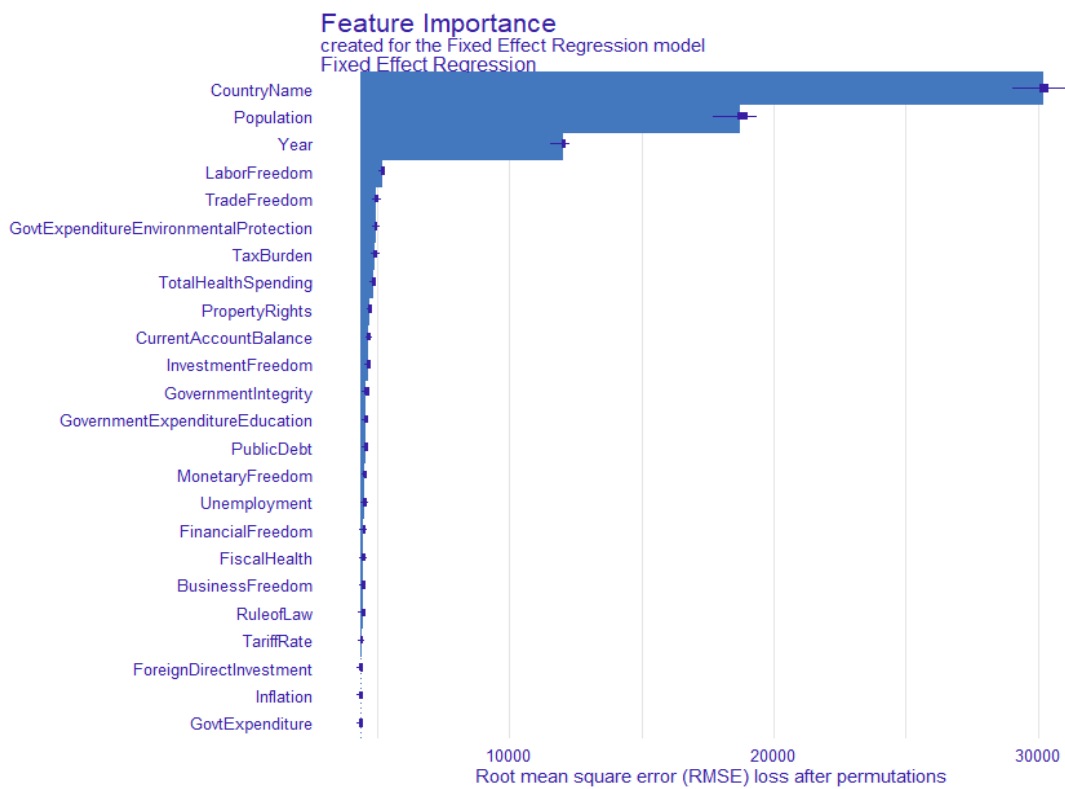
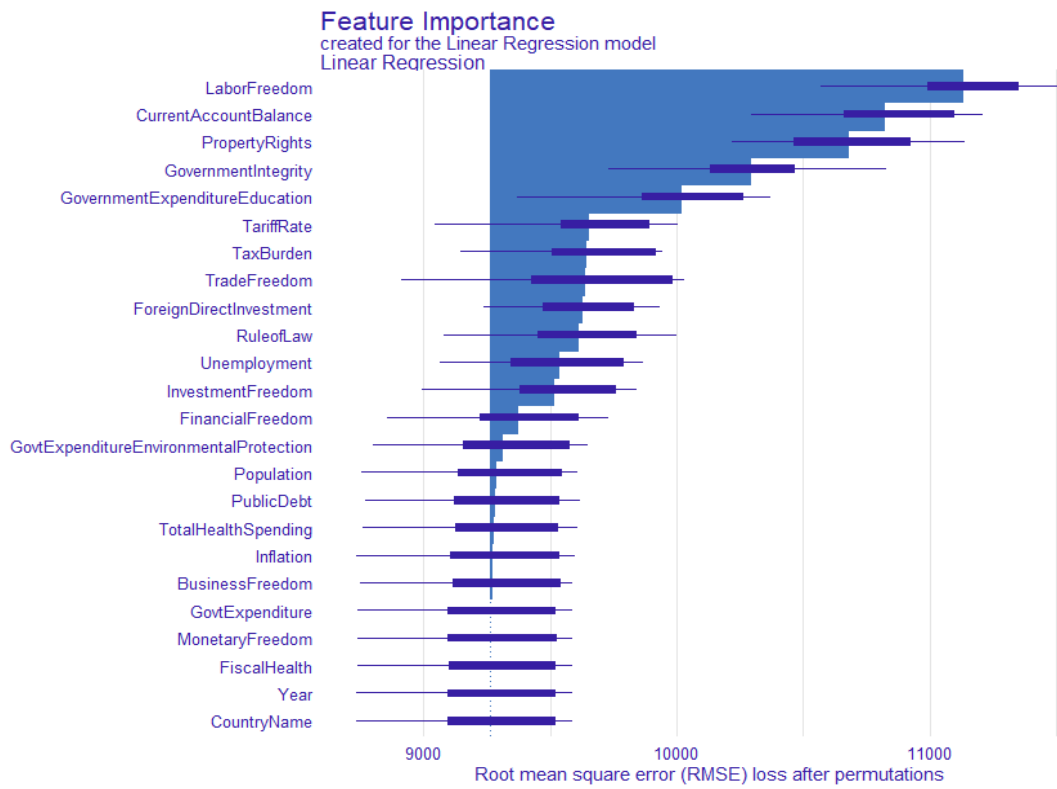


FIGURA A.11: Classical feature importance: POLS (esquerda) e Árvore de decisão (direita).



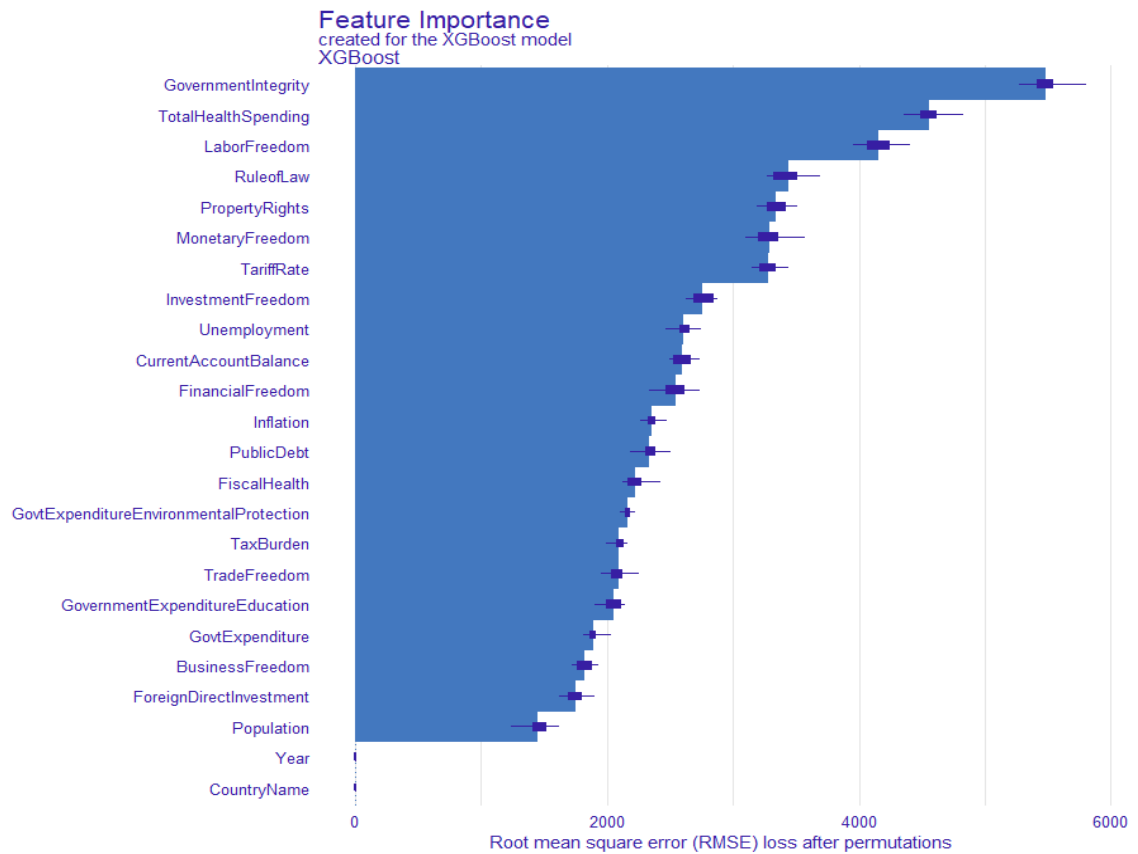
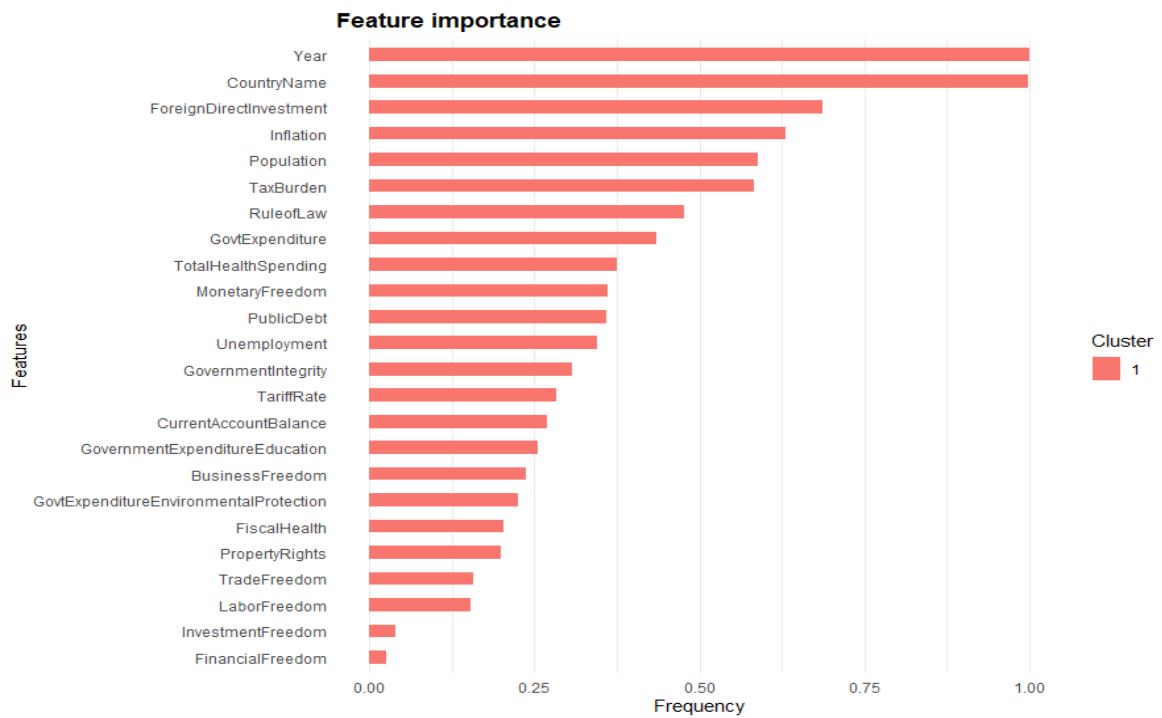


FIGURA A.12: *Permutation feature importance: POLS (cima (a)), Time-fixed Effects (meio (b)) e XGBoost (baixo (c)).*



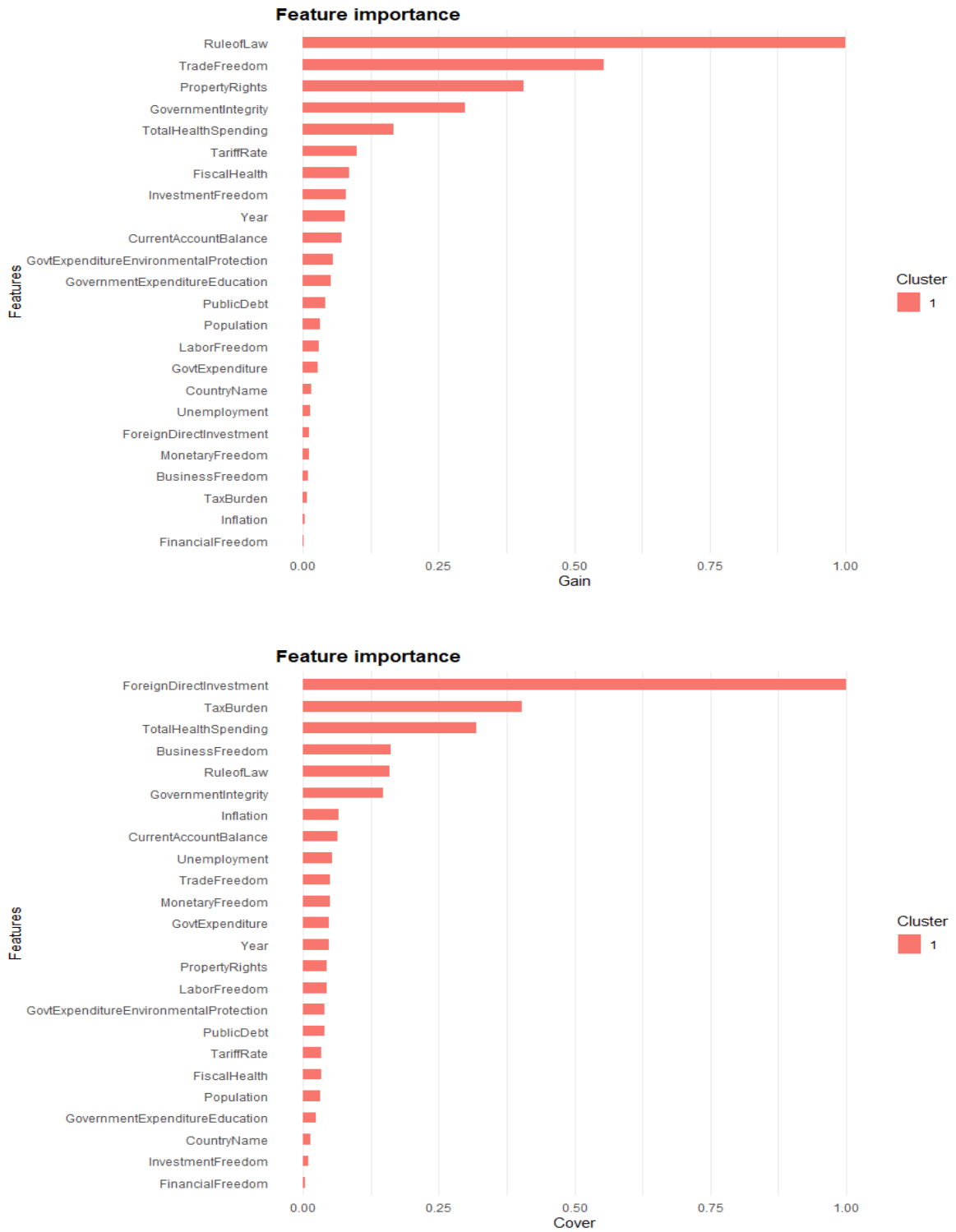
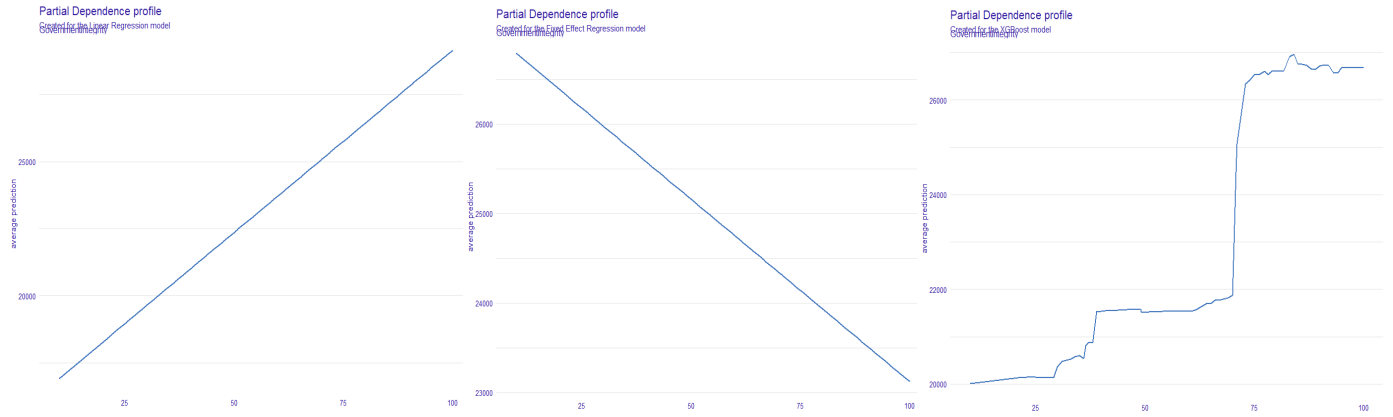
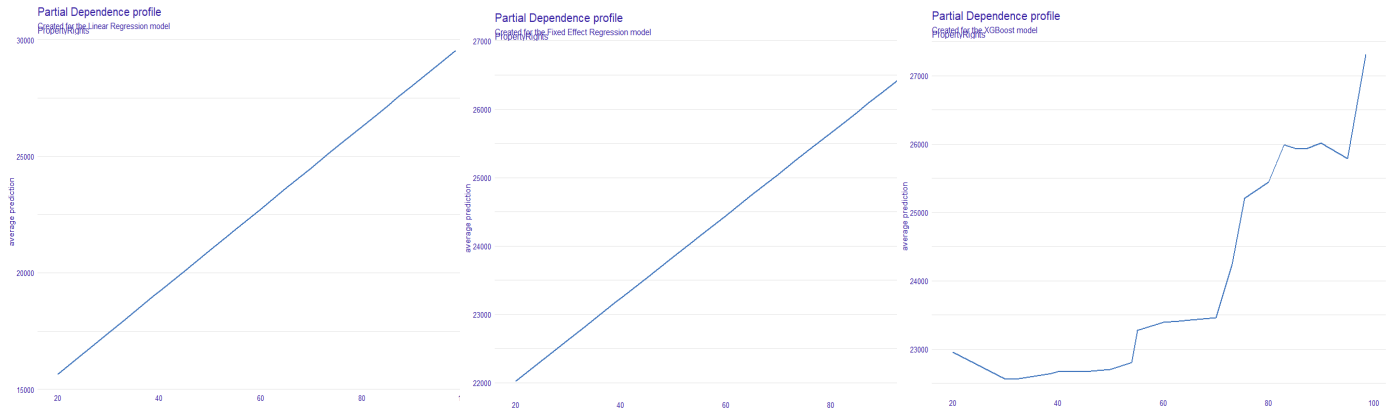


FIGURA A.13: *Feature importance* aplicado ao modelo XGBoost: frequência (cima (a)), ganho (centro (b)) e cobertura (baixo (c)).

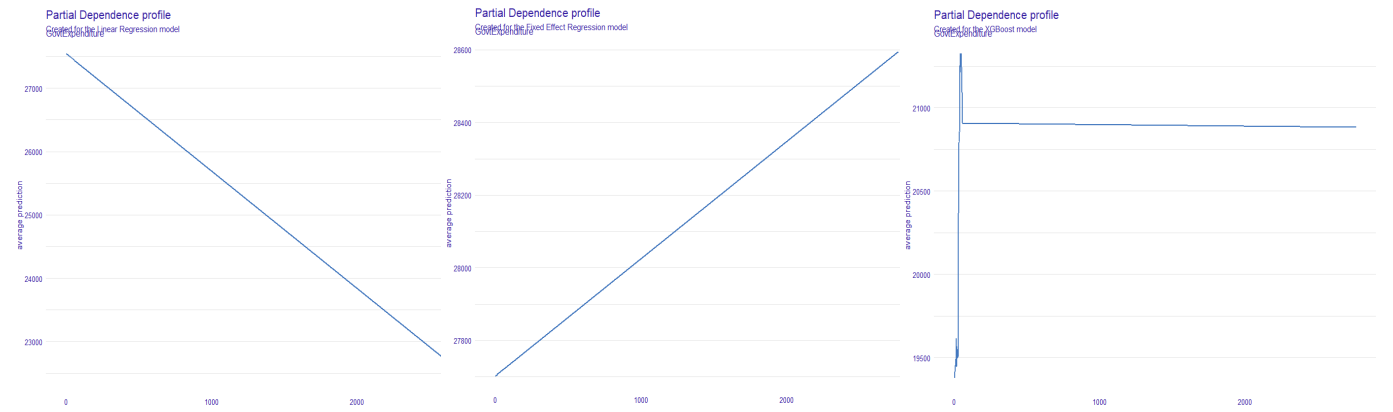
GovernmentIntegrity:



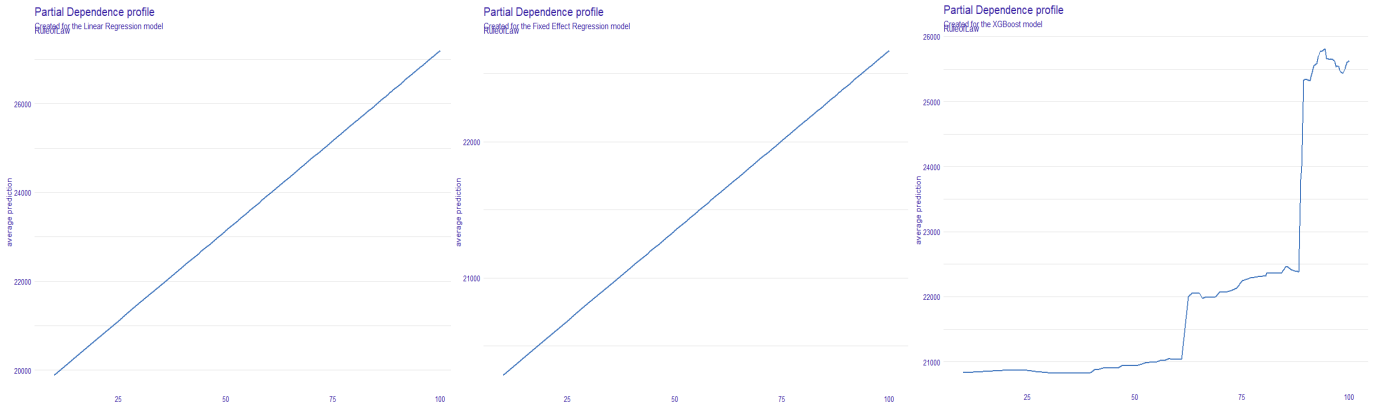
PropertyRights:



GovtExpenditure:



RuleofLaw:



TotalHealthSpending:

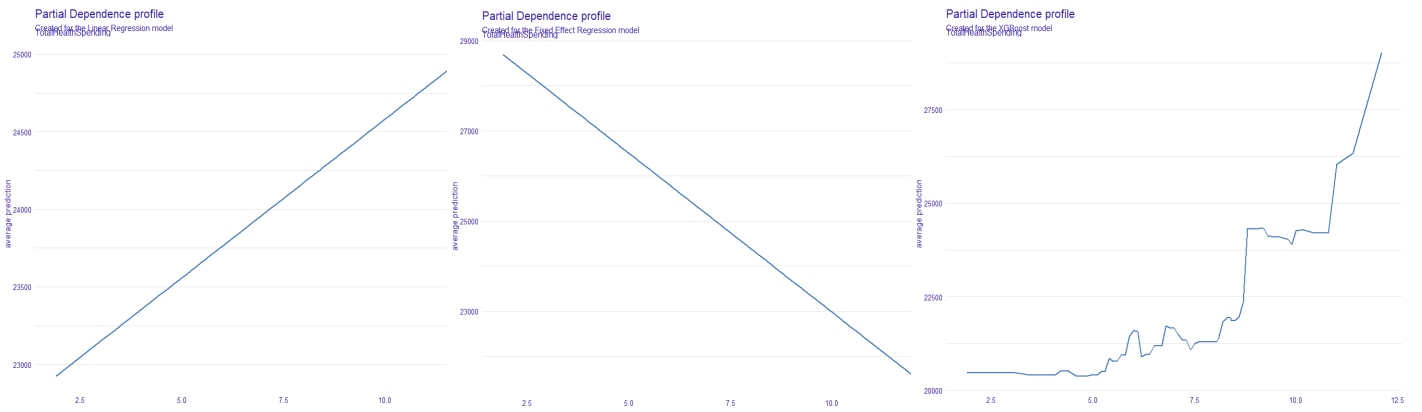


FIGURA A.14: Método PDP nos três modelos para as cinco variáveis mais importantes segundo o método SHAP values para XGBoost: PO (esquerda), TFE (centro) e XGB (direita).