Value Estimation of Properties Administered by the Brazilian Army Using Machine Learning and Spatial Components

José Nilo Alves de Sousa Neto[1], Marcelo Ladeira[1]

Universidade de Brasília
nilo.jose@aluno.unb.br, mladeira@unb.br

Abstract. The valuation of an institution's patrimony represents a necessary condition for an efficient management of its assets. The execution and analysis of real estate appraisal reports are essential to the achievement of some strategic objectives of the Brazilian Army, but they are also quite costly in terms of time, labor and financial resources. Sometimes, great effort is required for the aforementioned steps to take place and the market value finally obtained is inconsistent with what was initially imagined by the authorities, causing the technical study carried out to not be effectively used in negotiations by the organization. This work proposes the development of predictive models capable of building estimates of real estate values, so that the formal requests of the managers that imply the stages of execution and analysis of appraisal reports can occur with this information as an initial input. Counting on linear and nonlinear approaches and on machine learning techniques, the models have a reasonable level of assertiveness and national geographic coverage when generate estimated market values of Union real estate assets. Intrinsic and extrinsic variables to the properties were considered, including tests of aggregation of spatial components on some of them. As the interpretability of the proposed solution is an important requirement in both linear and nonlinear approaches, the Shapley value was adopted as a tool to support the guarantee of explainability and a PLS-SEM conceptual model was built to select attributes in a reasoned manner. These two considerations associated with modeling of real estate prices at a national level represent an innovation of this work in relation to the scientific literature analyzed.

CCS Concepts: • **Computing methodologies** → *Machine learning*.

Keywords: data mining, machine learning, real estate, spatial components

1. INTRODUÇÃO

Os bens tangíveis do tipo imobiliário possuem características intrínsecas e extrínsecas que os tornam únicos. Tal unicidade faz com que o processo de mensuração de seus valores mais prováveis e justos de mercado constitua matéria efetivamente complexa.

O Exército Brasileiro (EB) possui mais de vinte mil parcelas imobiliárias sob sua jurisdição e necessita de ferramenta para estimativa, em massa, dos valores desses bens patrimoniais a fim de suportar processos decisórios a nível estratégico visando à arrecadação de recursos para o Tesouro Nacional ou à permuta por outros ativos que melhor atendam suas atuais demandas.

No âmbito do EB, a Diretoria de Patrimônio Imobiliário e Meio Ambiente (DPIMA)[1] representa a instância técnica responsável por analisar e aprovar as avaliações imobiliárias dos imóveis administrados pelo EB e daqueles de interesse da Força Terrestre. Já a nível da União, a Secretaria de Coordenação e Governança do Patrimônio da União (SPU) faz a gestão dos ativos imobiliários.

Conforme a norma ABNT NBR 14653, o Método Comparativo Direto de Dados de Mercado deve ser preferencialmente utilizado; ele propõe a inferência dos valores de mercado dos imóveis com base

[1] http://www.dpima.eb.mil.br

na oferta ou na transação de outros imóveis com determinado nível de similaridade. A precificação de imóveis, ainda que de maneira estimativa, não é uma tarefa trivial. Trabalhos realizados mostram que a localização costuma se comportar como uma variável significante nos modelos construídos e que os imóveis ofertados ou recentemente transacionados nas adjacências daquele em análise têm influência sobre seu valor.

2. FUNDAMENTAÇÃO TEÓRICA

Segundo [Alves Dantas et al. 2010] e [Barros Antunes Campos and Almeida 2018], considerando modelo tradicional de preços hedônicos definido inicialmente por [Rosen 1974], o valor de mercado das unidades imobiliárias residenciais pode ser explicado por:

$$P = f(E, L, T, \beta) + \varepsilon \tag{1}$$

sendo o preço da habitação (P) função das suas características estruturais (E), locacionais (L) e da época em que foi demandado (T); f é um operador indicativo da forma funcional, β são parâmetros e ε os erros aleatórios do modelo.

Sob essa ótica, imóveis com outras vocações além da residencial também poderiam ter seus valores prováveis de mercado construídos a partir de características intrínsecas e extrínsecas.

3. TRABALHOS RELACIONADOS

Há diferentes abordagens com foco na valoração de imóveis presentes na literatura. [Alves Dantas et al. 2010] implementaram modelos de abrangência geográfica municipal combinados a econometria espacial. [Park and Bae 2015] aplicaram e analisaram algumas técnicas de aprendizagem de máquina, inclusive não lineares, no mercado de imóveis residenciais em Fairfax County, nos EUA. O algoritmo *Repeated Incremental Pruning to Produce Error Reduction* (RIPPER) apresentou a melhor performance geral, mas não houve enfoque em interpretabilidade.

[Kiely and Bastian 2020] desenvolveram modelos regressivos de predição de valores de imóveis de diferentes tipologias em Nova York com uso de redes neurais artificiais (ANN) e conceitos de agregação espacial de variáveis. [Dewan et al. 2019] combinaram ANN a modelos autorregressivos espaciais de caráter global (SAR) e os resultados obtidos indicaram a importância dos componentes espaciais.

[Hagenauer and Helbich 2022] propuseram um modelo de ANN combinado a regressões espaciais locais ponderadas, denominado *Geographically Weighted Artificial Neural Network* (GWANN). Quando aplicado ao domínio real de casas unifamiliares na Áustria, os maiores coeficientes médios de determinação (R^2) foram de aproximadamente 45%.

[Tchuente and Nyawa 2022] compararam o desempenho de 7 algoritmos de aprendizagem de máquina, lineares e não lineares, aplicados com e sem geocodificação a uma base de dados disponibilizada pelo Governo da França de valores de imóveis. Os modelos não lineares com geocodificação apresentaram maiores coeficientes de determinação, aproximadamente 74%.

Fundamentado em pesquisa bibliográfica realizada, a Tabela I apresenta um resumo comparativo entre a solução proposta e os trabalhos da literatura.

4. METODOLOGIA

Nesta seção, discorre-se acerca da extração de dados, de sua análise exploratória multivariada e da concepção inicial dos modelos associados à solução proposta.

Tabela I. Comparação com os trabalhos relacionados.

Trabalho	$R^2 > 70\%$	Abrangência nível País	Modelagem linear	Modelagem não linear	Ajuste de hiperparâmetros	Componentes espaciais	Mais de uma vocação de imóvel	Interpretabilidade
[Alves Dantas et al. 2010]	√		√			√		√
[Park and Bae 2015]	√		√	√	√			
[Kiely and Bastian 2020]			√	√	√	√	√	
[Hagenauer and Helbich 2022]		√	√	√		√		
[Tchuente and Nyawa 2022]	√	√	√	√	√	√		
Solução proposta	√	√	√	√	√	√	√	√

Tabela II. Informações sobre as instâncias rotuladas coletadas.

Fonte	Qtde instâncias coletadas	Perda geocodificação	Qtde pós-geocodificação	Participação relativa
EB	257	0 (0,0%)	257	6,0%
SPU	4981	925 (18,5%)	4056	94,0%
EB e SPU	5238	925 (17,6%)	4313	100,0%

4.1 Extração de Dados

Os dados diretamente associados a cada um dos imóveis foram extraídos de bases de dados de acesso restrito do EB e da SPU.

Todos os imóveis do EB foram georreferenciados em formato *shapefile*. Como as instâncias da SPU não continham polígonos georreferenciados, optou-se por geocodificar o campo textual *Endereço*. No processo de geocodificação, 925 instâncias não puderam ter suas coordenadas espaciais extraídas no *Google Earth Pro*, conforme resumo constante na Tabela II.

Essas bases foram enriquecidas com atributos socioeconômicos do censo 2010 do Instituto Brasileiro de Geografia e Estatística (IBGE) tratados por área de ponderação (AP). Para associar os imóveis às AP, realizaram-se operações de união espacial no *QGIS 3.16* entre os pontos relativos aos centroides das propriedades e os polígonos das AP, extraídos de [Furtado 2020]. As 4313 instâncias resultantes e as AP encontram-se representadas na Figura 1.

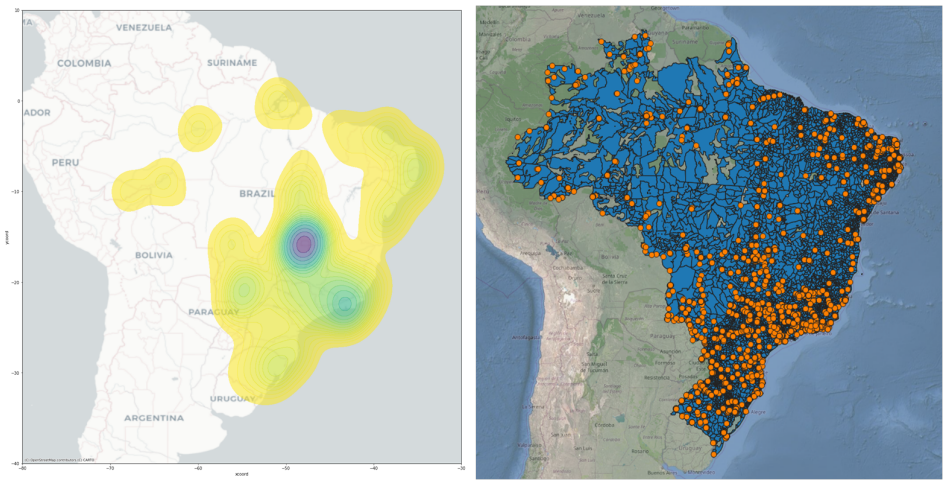

Fig. 1. Mapa de densidade dos imóveis (à esquerda) e AP sobrepostas pelos pontos (à direita). Fonte: autor.

Por fim, foram coletadas as quantidades de diversas tipologias de pontos de interesse, tais como estações de metrô, parques e hospitais, em um raio preferencial de 400 metros, tomando as coordenadas dos centroides dos imóveis como referência. Para tal, consumiu-se a API *Google Nearby Search Places*.

4.2 Tratamento de Instâncias

Foram desenvolvidas três abordagens para tratamento das instâncias coletadas, em decorrência da elevada dispersão observada das variáveis *Área do Terreno*, *Valor Total Atualizado* e *Valor Unitário Atualizado*, estas duas últimas atualizadas a janeiro de 2022 com uso do índice FipeZap Brasil[2] respeitando-se o tipo de uso do bem em questão. Segue uma síntese das abordagens abaixo:

(1) **Abordagem 1 ($\mathbf{A_1}$)** Remoção de *outliers* pelo método de Tukey baseada na variável *Valor Unitário Atualizado* (R\$/m^2) e utilização deste atributo como variável explicada.
(2) **Abordagem 2 ($\mathbf{A_2}$)** Transformação das variáveis *Área do Terreno* com logaritmo neperiano (ln), sem remover *outliers*, e ln de *Valor Total Atualizado* como variável explicada.
(3) **Abordagem 3 ($\mathbf{A_3}$)** Transformação do atributo *Valor Unitário Atualizado* com ln, remoção da variável *Área do Terreno* e ln de *Valor Unitário Atualizado* como variável explicada.

A abordagem A_2 apresentou melhores resultados, indicando necessidade de eliminação de apenas 2 instâncias das 4313 provenientes do processo de geocodificação. Na modelagem mais simples, linear, as 4311 ocorrências resultantes se mostraram não influenciantes sob avaliação da distância de Cook.

4.3 Análise e Seleção de Atributos

O critério para consideração inicial de um atributo foi baseado na sua disponibilidade sob forma estruturada ou passível de estruturação e pela sua presença nos bancos de dados do EB e da SPU.

A identificação dos subconjuntos de preditores úteis e alinhados ao pressuposto de não multicolinearidade exigido à aplicação de modelos de regressão linear múltipla foi realizada por meio de algumas técnicas específicas: análise de correlações bivariadas - critério C_1; regressão *forward-backward stepwise* (limites p-valor de entrada e de saída de 30%, fundamentados na NBR 14653) - critério C_2; eliminação recursiva de atributos com estimador regressivo *Random Forest* - critério C_3; e análise de significâncias com procedimento não paramétrico *bootstrapping* por meio da construção de modelo conceitual de equações estruturais (SEM, do inglês *structural equation modeling*) com mínimos quadrados parciais (PLS, do inglês *partial least squares*) - critério C_4. À exceção do PLS-SEM, implementado com uso do *software SmartPLS 4*, as demais técnicas foram implementadas com as bibliotecas *pandas* e *scikit-learn* em linguagem *Python*.

Segundo [Hair et al. 2022], resumido no *site* de apresentação do *SmartPLS*[3], "o PLS-SEM conta com um *bootstrapping* para testar a significância dos coeficientes de caminho estimados, semelhantes a regressores. No procedimento, subamostras são criadas com observações extraídas aleatoriamente do conjunto original de dados (com reposição). A subamostra é então usada para estimar o modelo de caminho PLS. Esse processo é repetido até que um grande número de subamostras aleatórias tenha sido criado, normalmente cerca de 10.000." Adotaram-se 10 mil neste trabalho.

A aplicação dos algoritmos de seleção de atributos resultou no que segue na Tabela III e no modelo conceitual PLS-SEM representado na Figura 2, na qual constam os modelos de medida (externos), caracterizados pelo relacionamento entre os indicadores em amarelo e as variáveis latentes exógenas, e o modelo estrutural (interno), marcado pelo relacionamento entre as variáveis latentes exógenas, de controle, as moderadoras e a endógena. Nos de medida, constam apenas os p-valores calculados

[2] `https://www.fipe.org.br/pt-br/indices/fipezap/`
[3] `https://www.smartpls.com/documentation/algorithms-and-techniques/bootstrapping/`

para as hipóteses de nulidade dos coeficientes de caminho com base em distribuição t de Student; já para o estrutural, constam os coeficientes de caminho de caminho em si, os p-valores calculados e o coeficiente de determinação para a variável latente endógena *Ln(Valor Total Atualizado)*.

Analisando-se a relação entre o constructo formativo *Caracterização do Terreno e das Benfeitorias* e a variável latente endógena *Valor do Imóvel* ilustrados na Figura 2, percebe-se que se trata da conexão mais forte do modelo estrutural.

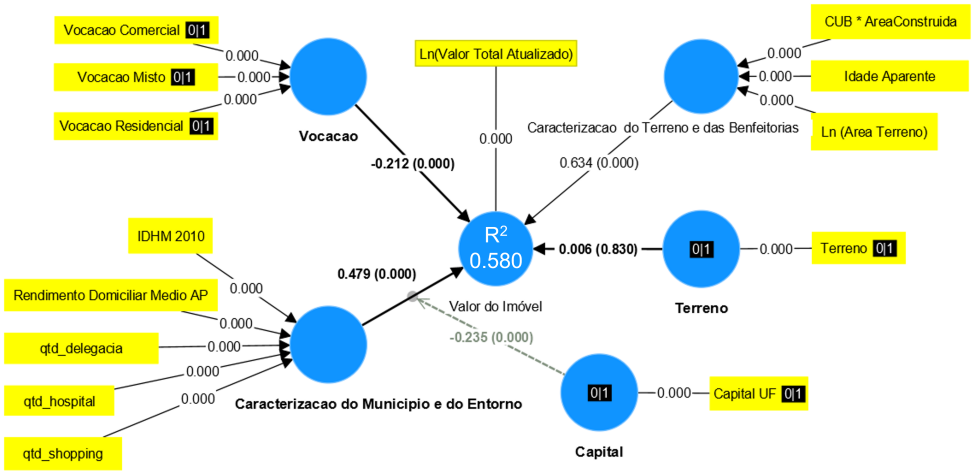

Fig. 2. Modelo conceitual PLS-SEM construído no *software SmartPLS 4*.

Tabela III. Ocorrência de atributos nas modelagens finais.

Variável	**Tipo**	**Fonte**	**Critério de Exclusão**
Capital UF	binária	IBGE	-
Tipologia Municipal	cód. binários	IBGE	C_2
Grau de Urbanização 2010	numérica	IBGE	C_4
Índice de Desenvolvimento Humano 2010	numérica	IBGE	-
Índice de Vulnerabilidade Social 2010	numérica	Ipea	C_4
Taxa de Homicídios 2019	numérica	Ipea	C_4
Vocação do Imóvel	cód. binários	EB/SPU	-
Terreno	binária	EB/SPU	-
Ln(Área do Terreno)	numérica	EB/SPU	-
*CUB * Área Construída*	numérica	Sinduscon[a]/EB/SPU	-
Idade Aparente	cód. alocados	EB/SPU	-
Vida Útil	numérica	BIR[b]	C_4
% Superior Completo AP 2010	numérica	IBGE	C_4
% Rede Geral de Distribuição de Água AP 2010	numérica	IBGE	C_4
% Microcomputador com Acesso à Internet AP 2010	numérica	IBGE	C_1
Rendimento Domiciliar Médio AP 2010	numérica	IBGE	-
Pontos de Interesse[c]	numérica	API *Google*	-
Pontos de Interesse Excluídos[d]	numérica	API *Google*	C_4
Coordenadas Geográficas[e]	numérica	EB/SPU	C_3
Ln(Valor Total Atualizado)	numérica	EB/SPU	-

[a] https://www.cub.org.br/
[b] *Bureau of Internal Revenue*
[c] delegacias, hospitais e *shopping centers*
[d] parques, escolas, universidades, atrações turísticas, supermercados, restaurantes e estações de metrô
[e] utilizadas para uniões espaciais, consultas API *Google* e construção de matriz de pesos espaciais.

4.4 Tratamento de Variáveis

Foram aplicados testes de normalidade numéricos aos atributos selecionados. Tanto o teste de Shapiro-Wilk quanto o de Jarque-Bera indicaram, pelos p-valores obtidos, que há evidências de que os dados têm assimetria e curtose significativamente diferentes de uma distribuição normal.

Optou-se, portanto, pela aplicação de transformação do tipo *MinMaxScaler*, utilizando a biblioteca *scikit-learn* em *Python*. Tal etapa evita que variáveis com maior amplitude de variação influenciem demasiadamente o modelo, explicando desproporcionalmente a variância do atributo dependente.

4.5 Modelos Preditivos

À luz das referências bibliográficas coletadas, configuraram-se e testaram-se os seguintes algoritmos de predição: regressão linear múltipla (OLS), regressão espacial (GM_Lag), SGDRegressor linear (aprendizagem de máquina) e XGBoost não linear (aprendizagem de máquina). As bibliotecas em linguagem *Python* utilizadas foram: *statsmodel*, *scikit-learn* e *spreg*.

O modelo de regressão espacial teve como fundamento principal a defasagem espacial da variável explicada *Ln(Valor Total Atualizado)* associada a método dos mínimos quadrados espaciais de 2 estágios (S2SLS), conforme proposto por [Anselin 1988]. A matriz de pesos espacias W foi construída com função de decaimento do inverso da distância euclidiana a cada imóvel rotulado da base de dados até um limite de raio de influência de 1 quilômetro. Foram testadas outras funções e raios de influência e os adotados foram os que apresentaram resultados mais robustos.

Na documentação[4] de gradiente descendente estocástico (SGD), é possível verificar que ele é uma rotina simples, mas bastante eficiente quanto ao ajuste de regressores lineares sob funções de perda convexa. A rigor, SGD representa uma técnica de otimização e uma maneira de treinar modelos. Já o XGBoost é um algoritmo projetado para ser altamente eficiente, flexível e portátil, com funções de custo e de regularização bastante adaptáveis. Sua formulação matemática pode ser consultada no *site*[5] da biblioteca.

5. EXPERIMENTOS E RESULTADOS

Nesta seção, são apresentadas as configurações dos modelos de aprendizagem de máquina e os resultados obtidos, considerando-se as bases de dados concatenadas e tratadas do EB e da SPU enriquecidas com informações do IBGE e das consultas ao *Google Places*.

5.1 Configuração dos Experimentos

A configuração dos principais hiperparâmetros dos modelos de aprendizagem de máquina contida na Tabela IV foi feita de forma minuciosa, considerando várias combinações possíveis de taxas de aprendizagem, sementes aleatórias de inicialização de pesos, termos de regularização e profundidade dos estimadores adotados. A partição entre os conjuntos de treinamento, validação e teste foi feita de forma estratificada, com base na vocação dos imóveis.

5.2 Métricas Avaliadas

O coeficiente de determinação (R^2) foi a principal métrica avaliada, respeitando-se limites de tempo de processamento razoáveis. Uma outra métrica analisada foi a raiz quadrada da média dos erros (RMSE). As médias e os desvios padrões do R^2 foram calculados para 10 *folds* em validação cruzada.

[4] `https://scikit-learn.org/stable/modules/sgd`
[5] `https://xgboost.readthedocs.io/en/stable/index.html`

Tabela IV. Parâmetros utilizados nos modelos de aprendizagem de máquina.

Parâmetro	**SGDRegressor linear**	**XGBoost não linear**
Fração de Treinamento	53,3%	53,3%
Fração de Validação	13,3%	13,3%
Fração de Teste	33,3%	33,3%
Tipo de taxa de aprendizagem	constante	constante
Valor da taxa de aprendizagem	0,0001	0,5
Função de custo	erro quadrático	erro quadrático
Termo de regularização	l2	l2
Profundidade máxima de uma árvore	não se aplica	5

Tabela V. Resultados de coeficiente de determinação para os dois cenários considerados.

Modelo	**R^2 Médio**	**dp dos R^2 para 10 *folds* Treinamento**	**R^2 Teste**	**RMSE**
Regressão Linear Múltipla	61,58%	não se aplica	não se aplica	1,29
Regressão Espacial	63,08%	não se aplica	não se aplica	1,35
SGDRegressor linear (cenário A)	60,61%	0,042	59,00%	1,30
SGDRegressor linear (cenário B)	60,11%	0,027	45,56%	1,38
XGBoost não linear (cenário A)	83,21%	0,031	82,74%	0,84
XGBoost não linear (cenário B)	84,65%	0,025	80,84%	0,82

5.3 Resultados Obtidos

Foram obtidos os resultados especificados na Tabela V para os modelos lineares e não lineares implementados em ambiente *Google Research Colaboratory* (Colab) com uso de linguagem *Python*.

A aplicação dos algoritmos de aprendizagem de máquina foi realizada em 2 cenários: A (treinamento, validação e teste em base de dados híbrida do EB e da SPU) e B (treinamento e validação em base híbrida e teste em base exclusiva do EB composta por 130 instâncias, aproximadamente 50% do total de ocorrências rotuladas de imóveis administrados pelo EB).

Os R^2 e as medidas de dispersão apresentados na Tabela V indicam predições mais assertivas do modelo não linear. Vale ressaltar que a modelagem com XGBoost apresentou melhor performance com utilização de mais variáveis explicativas, entretanto, optou-se por manter todos os modelos construídos na mesma condição, a fim de compará-los sob estados semelhantes.

Para o modelo de regressão espacial, analisou-se a dependência espacial, autocorrelação global, a partir do índice de Moran e do teste de Anselin-Kelejian. Ambos indicaram que os componentes espaciais são significantes.

5.4 Interpretabilidade dos Resultados Obtidos

A interpretabilidade utilizando o valor de Shapley pode ser garantida por meio de definições colaborativas da Teoria dos Jogos. Conforme [Lundberg et al. 2020], o valor de Shapley é a contribuição marginal média de cada valor de atributo em todas as combinações possíveis de características. Os atributos com grandes valores de Shapley absolutos são considerados importantes.

As variáveis mais importantes em cada modelo de aprendizagem de máquina estão posicionadas na porção superior dos gráficos ilustrados na Figura 3. Verifica-se que a relação das variáveis explicativas com a variável dependente se mostra coerente com a realidade observada no domínio em estudo. A variável *Ln(Área do Terreno)* se mostra como o atributo mais importante na formação de valor dos imóveis tanto na abordagem linear quanto na não linear. As características locacionais, *IDHM* a nível de granularidade espacial municipal e *Rendimento Domiciliar Médio* a nível AP, se mostram bem mais importantes na modelagem não linear.

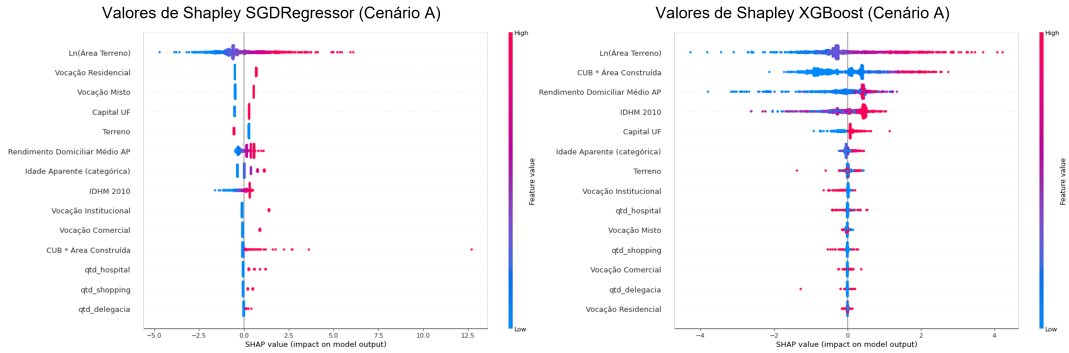

Fig. 3. Valores de Shapley calculados para os modelos SGDRegressor (à esquerda) e XGBoost (à direita).

6. CONCLUSÃO E TRABALHOS FUTUROS

Este trabalho contribui em vertente de inovação ao EB, mais especificamente nas áreas de valoração imobiliária e de computação aplicada à resolução de problemas reais. A documentação de metodologia clara, respeitando os preceitos de conhecimento de mineração de dados e de estatística quanto à dinâmica de precificação de imóveis, corrobora para o desenvolvimento de uma modelagem inovadora.

Atingiram-se coeficientes de determinação razoáveis, em relação aos trabalhos referenciados, na tentativa de obter modelos capazes de construir estimativas de valores de imóveis administrados pelo EB e pela SPU nacionalmente. Complementarmente, o PLS-SEM se mostrou interessante ao entendimento dos vínculos entre os atributos e à sua seleção; sua combinação a modelos de aprendizagem de máquina visando à precificação de imóveis representa uma contribuição deste trabalho. É válido ressaltar que há limitações quanto à disponibilidade de dados de mercado avaliados à luz da NBR 14653 ou efetivamente transacionados. Como passos seguintes, pretende-se explorar modelos com uma quantidade mais robusta de variáveis, possivelmente, incluindo informações do censo IBGE 2022.

REFERENCES

Alves Dantas, R., Magalhães, A. M., and Vergolino, J. R. d. O. Um Modelo Espacial de Demanda Habitacional para a Cidade do Recife. *Estudos Econômicos (São Paulo)* 40 (4): 891–916, 2010.

Anselin, L. *Spatial Econometrics: Methods and Models*. Springer Dordrecht, 1988.

Barros Antunes Campos, R. and Almeida, E. Decomposição espacial nos preços residenciais no município de São Paulo. *Estudos Econômicos (São Paulo)* 48 (1): 5–38, 2018.

Dewan, P., Ganti, R., Srivatsa, M., and Stein, S. NN-SAR: A Neural Network Approach for Spatial AutoRegression. In *2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. Kyoto, Japan, pp. 783–789, 2019.

Furtado, B. A. *NT DISET 78 - Gerando Famílias Artificiais Intraurbanas: censo 2010*. Ipea, 2020.

Hagenauer, J. and Helbich, M. A geographically weighted artificial neural network. *International Journal of Geographical Information Science* 36 (2): 215–235, 2022.

Hair, J., Hult, G. T. M., Ringle, C., and Sarstedt, M. *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)*. SAGE Publications, Inc, 2022.

Kiely, T. and Bastian, N. The spatially conscious machine learning model. *Statistical Analysis and Data Mining: The ASA Data Science Journal* 13 (1): 31–49, 2020.

Lundberg, S., Erion, G., Chen, H., DeGrave, A., Prutkin, J., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* vol. 2, pp. 56–67, 2020.

Park, B. and Bae, J. Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Systems with Applications* 42 (6): 2928–2934, 2015.

Rosen, S. Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy* 82 (1): 34–55, 1974.

Tchuente, D. and Nyawa, S. Real estate price estimation in French cities using geocoding and machine learning. *Annals of Operations Research* 308 (1-2, SI): 571–608, 2022.