Ciência de dados aplicada às eleições municipais: análise preditiva do sucesso eleitoral em Rondônia

Fábio Augusto Ferreira[1]*; Thiago Gentil Ramires[2]

[1] Governo do Estado de Rondônia. Auditor Fiscal de Tributos Estaduais. Avenida Luiz Mazieiro, 4060 – Jardim América; 76.980-726 Vilhena, Rondônia, Brasil.
[2] Universidade Tecnológica Federal do Paraná. Diretor de Pesquisa e Pós-Graduação. Rua Marcílio Dias, 635 – Jardim Paraíso; 86812460 Apucarana, PR - Brasil
*autor correspondente: fabioferreira@sefin.ro.gov.br

Ciência de dados aplicada às eleições municipais: análise preditiva do sucesso eleitoral em Rondônia

Resumo

A análise preditiva do sucesso eleitoral representa uma fronteira inovadora na aplicação de técnicas de Data Science ao campo da ciência política brasileira. Este trabalho objetivou desenvolver e comparar modelos preditivos para o resultado eleitoral municipal em Rondônia, utilizando técnicas de machine learning sobre dados estruturados de financiamento de campanha, perfil de candidatos e características municipais. Foram utilizados dados do Tribunal Superior Eleitoral (TSE) e do Instituto Brasileiro de Geografia e Estatística (IBGE) referentes aos pleitos de 2020 e 2024. Foram implementados seis algoritmos de classificação, incluindo Regressão Logística, “Random Forest”, “Gradient Boosting” e métodos de “ensemble”, avaliados por validação cruzada e teste temporal (treino 2020; teste 2024). Os resultados indicaram que o modelo “Gradient Boosting” apresentou a melhor performance, alcançando “F1-Score” de 0,2836 e ROC-AUC de 0,5028. A análise de importância de variáveis revelou a dominância da variável total_receita, seguida por gestão administrativa e materiais gráficos e sonoros, evidenciando que tanto o volume de recursos quanto a eficiência na gestão influenciam os resultados eleitorais. A aplicação de técnicas de interpretabilidade (SHAP e LIME) confirmou os padrões globais e identificou perfis distintos de candidatos, demonstrando que a boa gestão pode compensar restrições orçamentárias. Conclui-se que o sucesso eleitoral é multifatorial, com forte componente de alocação de recursos, e que o “framework” proposto contribui para a transparência democrática e para a formulação de estratégias de campanha baseadas em evidências.
Palavras-chave: aprendizado de máquina; financiamento de campanhas; interpretabilidade; SHAP; LIME.

Data Science Applied to Municipal Elections: Predictive Analysis of Electoral Success in Rondônia

Abstract

Predictive analysis of electoral success represents an innovative frontier in the application of Data Science techniques to the field of Brazilian political science. This work aimed to develop and compare predictive models for municipal electoral outcomes in Rondônia, using machine learning techniques on structured data from campaign financing, candidate profiles, and municipal characteristics. Data from the Brazilian Electoral Court (TSE) and the Brazilian Institute of Geography and Statistics (IBGE) covering the 2020 and 2024 elections. Six classification algorithms were implemented, including Logistic Regression, Random Forest, “Gradient Boosting”, and ensemble methods, evaluated through cross-validation and temporal testing (training on 2020; testing on 2024). Results showed that “Gradient Boosting” achieved the best performance, with an “F1-Score” of 0.2836 and ROC-AUC of 0.5028. Feature importance analysis revealed the dominance of total_receita as the main predictor, followed by administrative management and graphic and sound materials, indicating that both resource volume and management efficiency are critical for electoral success. The application of explainability methods (SHAP and LIME) confirmed global patterns and highlighted distinct candidate profiles, showing that efficient management can offset budget limitations. The findings suggest that electoral success is multifactorial, with a strong component of resource allocation efficiency. The proposed “framework” contributes to electoral transparency and offers actionable insights for evidence-based campaign strategies in Brazilian municipalities.
Keywords: machine learning; campaign financing; interpretability; SHAP; LIME.

Introdução

O processo eleitoral brasileiro constitui um fenômeno complexo e multifacetado, em que variáveis financeiras, sociais, políticas e demográficas interagem para determinar os resultados. O estado de Rondônia, localizado na Região Norte do Brasil, apresenta particularidades que o tornam um campo de pesquisa especialmente fértil. Seus 52 municípios exibem grande diversidade socioeconômica, com IDHM variando entre 0,600 e 0,750. Suas economias locais ancoram-se na agropecuária, extrativismo e comércio, e a estrutura partidária é fragmentada, típica do cenário político brasileiro.

Entre 2020 e 2024, observou-se uma transformação relevante nesse ambiente: a introdução e consolidação do Fundo Especial de Financiamento de Campanha (FEFC), que alterou profundamente as estratégias de arrecadação e alocação de recursos eleitorais, gerando um cenário em que a análise quantitativa das campanhas se torna ainda mais necessária para compreensão dos determinantes do sucesso eleitoral.

Nesse contexto, a análise preditiva baseada em ciência de dados surge como uma ferramenta de grande potencial. O avanço dos métodos de “machine learning” permite não apenas identificar padrões complexos em grandes volumes de dados, mas também gerar previsões com graus crescentes de confiabilidade.

Entretanto, a mera previsão do resultado não é suficiente no campo da ciência política aplicado, já que o desafio consiste em conciliar capacidade preditiva com interpretabilidade, de tal modo que os modelos desenvolvidos forneçam explicações claras sobre quais variáveis influenciam o sucesso eleitoral e em que magnitude e sob quais condições.

Embora a literatura internacional sobre financiamento de campanhas seja vasta, destacando reiteradamente a relação positiva entre volume de recursos e desempenho eleitoral (Bartels, 2008; Samuels, 2001), ainda existem lacunas importantes no Brasil, especialmente quando se trata de estudos aplicados a contextos regionais menos explorados, como os estados da Região Norte.

Além disso, mesmo quando modelos preditivos são aplicados ao processo eleitoral, observa-se uma distância entre a capacidade de previsão e a possibilidade de interpretação. Modelos sofisticados, como “Gradient Boosting” e “Random Forest”, oferecem desempenho elevado, mas frequentemente falham em oferecer clareza sobre os fatores determinantes de suas decisões. Isso cria um déficit explicativo que limita a aplicabilidade prática das descobertas para candidatos, estrategistas e gestores públicos, que necessitam de previsões acompanhadas de explicações compreensíveis.

Dessa forma, o problema central desta pesquisa pode ser formulado nos seguintes termos: como desenvolver um “framework” preditivo que seja simultaneamente robusto em

termos de performance e transparente em termos de interpretabilidade, aplicado às eleições municipais de Rondônia nos ciclos de 2020 e 2024? A resposta a esse problema envolve o emprego de técnicas avançadas de ciência de dados, não apenas para a construção de modelos preditivos, mas também para explicar, de forma global e local, quais fatores impulsionam as chances de sucesso de um candidato.

Assim, o objetivo é desenvolver e validar um "framework" preditivo interpretável para análise de determinantes do sucesso eleitoral em municípios de Rondônia, integrando algoritmos de "machine learning" com métodos de interpretabilidade de modelos ("explainable AI"), especificamente ao: I - Construir e consolidar uma base de dados multidimensional abrangendo variáveis financeiras, socioeconômicas, eleitorais e administrativas; II - Implementar técnicas de "feature" "engineering" para reduzir multicolinearidade e agrupar variáveis em dimensões significativas, validando o processo por meio de correlação de Pearson e "Variance Inflation Factor" (VIF); III - Treinar, otimizar e comparar o desempenho de seis algoritmos de "machine learning" (SVM, "Random Forest", "Gradient Boosting", "Voting", "Bagging" e "Stacking"), considerando métricas adequadas a problemas de classificação desbalanceada, como "F1-Score" e ROC-AUC. IV - Aplicar métodos de interpretabilidade global (SHAP) e local (LIME) para identificar variáveis críticas, perfis eleitorais típicos e caminhos alternativos para o sucesso eleitoral. V - Avaliar a robustez e estabilidade temporal dos modelos, validando-os por meio de divisão temporal (treino: 2020; teste: 2024) e validação cruzada estratificada. VI - Traduzir os achados técnicos em insights estratégicos aplicáveis a campanhas políticas, contribuindo para a transparência democrática e para a literatura de ciência política aplicada.

Esta pesquisa justifica-se porque conecta ciência de dados e ciência política, em um contexto ainda pouco explorado no Brasil, oferecendo evidências empíricas inéditas e aplicabilidade prática imediata. Na prática, o estudo oferece instrumentos concretos para a formulação de estratégias de campanha mais eficientes e transparentes, permitindo que candidatos, partidos e gestores públicos baseiem suas decisões em evidências quantitativas e explicáveis.

Além disso, os resultados contribuem para o fortalecimento da democracia ao ampliar a compreensão sobre o papel do financiamento eleitoral e da gestão de campanha, reforçando a importância da transparência e da equidade no processo político.

Portanto, este trabalho tem como objetivo desenvolver um "framework" preditivo interpretável, com base em técnicas de "machine learning" e "explainable AI", capaz de identificar os determinantes do sucesso eleitoral municipal em Rondônia nos ciclos de 2020 e 2024.

Metodologia

A escolha metodológica foi pautada pela necessidade de aliar robustez estatística, adequação ao objetivo de pesquisa e transparência interpretativa, garantindo que os resultados obtidos pudessem não apenas prever o sucesso eleitoral com confiabilidade, mas também explicar os fatores subjacentes que o determinam. O desenho metodológico, portanto, foi estruturado em cinco dimensões principais: coleta e preparação dos dados, engenharia e validação das variáveis, modelagem preditiva com algoritmos de “machine learning”, implementação de técnicas de interpretabilidade (“explainable AI” – XAI) e avaliação da performance com validação temporal.

Por fim, os dados e *scripts* para geração dos modelos da pesquisa estão disponibilizados em repositório: https://github.com/FabioAFerreira/TCC.DSA.USP.ESALQ.

Coleta e Preparação dos Dados

O primeiro passo consistiu na seleção e consolidação das bases de dados necessárias para o estudo. Foram utilizados dados secundários de domínio público, provenientes principalmente do Tribunal Superior Eleitoral (TSE), que disponibiliza informações detalhadas sobre candidaturas, receitas e despesas de campanha, resultados eleitorais e situação jurídica dos candidatos.

Complementarmente, o Instituto Brasileiro de Geografia e Estatística (IBGE), forneceu indicadores socioeconômicos e demográficos dos municípios de Rondônia, incluindo Produto Interno Bruto (PIB) per capita, Índice de Desenvolvimento Humano Municipal (IDHM), densidade demográfica, taxa de alfabetização e percentual de população urbana, extraídos do Censo 2022 (IBGE, 2023). O período de análise contemplou dois ciclos eleitorais completos, correspondentes aos anos de 2020 e 2024, de forma a capturar não apenas padrões estáticos, mas também a evolução temporal das dinâmicas de financiamento e resultado eleitoral.

A preparação dos dados envolveu um pipeline rigoroso de “Extract”, “Transform”, “Load” (ETL), desenvolvido em “Python” 3.9 e estruturado com o auxílio das bibliotecas “Pandas” e “NumPy”. A etapa de extração envolveu a importação dos arquivos em formato CSV disponibilizados pelo TSE e IBGE, com detecção automática de “encoding” para assegurar a preservação de caracteres especiais e acentuação própria da língua portuguesa.

Na etapa de transformação, procedeu-se à padronização dos formatos de variáveis, à normalização de valores monetários para “floats” numéricos e à unificação de nomenclaturas de municípios e cargos políticos. A limpeza de dados incluiu a identificação e imputação de

valores faltantes por meio de métodos multivariados, considerando a similaridade entre municípios com base em indicadores socioeconômicos, e a exclusão de registros inconsistentes ou duplicados. Por fim, a etapa de “load” consistiu na consolidação de um “dataset” final com 18 variáveis preditivas estrategicamente agrupadas e uma variável alvo binária (eleito = 1; não eleito = 0), conforme detalhado na Tabela 1.

Tabela 1. Agrupamento de variáveis

Grupo	Variável	Descrição	Fonte
Financeira	total_receita	Receita total declarada da campanha	TSE
	gestao_administrativa	Despesas com serviços contábeis, jurídicos e gestão organizacional	TSE
	midia_propaganda	Gastos em rádio, TV, impulsionamento digital e redes sociais	TSE
	materiais_graficos_sonoros	Despesas com impressos, adesivos e carros de som	TSE
	mobilizacao_humana	Custos de contratação de pessoal e militância de rua	TSE
	infraestrutura_basica	Custos de suporte logístico e físico (alimentação, imóveis, transporte)	TSE
	apoio_politico	Despesas com eventos e apoio a candidatos/partidos	TSE
	pesquisa_eleitoral	Recursos investidos em pesquisas e sondagens eleitorais	TSE
	aquisicoes_bens	Gastos com bens móveis, veículos e equipamentos de campanha	TSE
	comunicacao_correspondencia	Despesas com correspondências e serviços postais	TSE
Socioeconômica	idhm	Índice de Desenvolvimento Humano Municipal	IBGE
	pib_per_capita	PIB per capita do município	IBGE
	habitantes	População residente total do município	IBGE
	total_receitas_brutas_ibge	Receita bruta municipal (indicador fiscal)	IBGE
Eleitoral/Política	NR_PARTIDO	Sigla do partido do candidato	TSE
	eleitores	Número total de eleitores aptos no município	TSE
	vagas	Número de cargos em disputa (prefeito/vereadores)	TSE
	diversos	Outras despesas não classificadas nas categorias principais	TSE

Fonte: Dados originais da pesquisa

Despesas eleitorais

A análise detalhada das despesas eleitorais constitui um passo essencial para compreender a lógica subjacente ao financiamento das campanhas municipais em Rondônia, além de fornecer subsídios diretos para a construção das variáveis agregadas utilizadas na modelagem preditiva.

Entre os ciclos eleitorais de 2020 e 2024, verificou-se um crescimento expressivo do volume total de recursos movimentados, passando de R$ 78,7 milhões para R$ 110,7 milhões, o que representa uma expansão de 40,66% em valores absolutos, conforme descrito na Tabela 2.

Tabela 2. Despesas Eleitorais

Categoria	2020 (R$ mi)	2020%	2024 (R$ mi)	2024%	Variação Absoluta (R$ mi)	Crescimento (%)
Gestão Administrativa	21	26,70%	28,5	25,80%	7,5	35,70%
Mobilização Humana	14,2	18,00%	19,8	17,90%	5,6	39,40%
Mídia e Propaganda	12,5	15,90%	18,5	16,70%	6	48,00%
Materiais Gráficos/Sonoros	8,9	11,30%	12,4	11,20%	3,5	39,30%
Infraestrutura Básica	8	10,20%	11,2	10,10%	3,2	40,00%
Apoio Político	4	5,10%	6	5,40%	2	50,00%
Pesquisas Eleitorais	2,7	3,40%	4,8	4,30%	2,1	77,80%
Aquisições de Bens	3,6	4,60%	5,4	4,90%	1,8	51,10%
Comunicação/Correspondência	2,1	2,70%	3,4	3,10%	1,3	61,90%
Diversos	1,7	2,10%	3	2,70%	1,3	75,00%
TOTAL	**78,7**	**100%**	**110,7**	**100%**	**32**	**40,70%**

Fonte: Dados originais da pesquisa

Este aumento ocorreu em um contexto de redução do número de candidatos de 104 para 98 (queda de 5,77%), o que sugere não apenas maior seletividade no processo eleitoral, mas também uma concentração mais intensa de recursos por candidatura.

Em termos médios, o investimento por candidato cresceu aproximadamente 49,25%, evidenciando um processo de profissionalização e sofisticação crescente das estratégias de campanha.

A distribuição das despesas por categoria revela ainda transformações estruturais relevantes no modo como os recursos eleitorais vêm sendo alocados. Em 2024, destacaram-se como categorias dominantes: gestão administrativa (25,75%; R$ 28,5 milhões), mobilização humana (17,89%; R$ 19,8 milhões), mídia e propaganda (16,71%; R$ 18,5 milhões), materiais gráficos e sonoros (11,20%; R$ 12,4 milhões) e infraestrutura básica

(10,12%; R$ 11,2 milhões). Essas cinco categorias somadas concentraram mais de 80% das despesas totais, corroborando a hipótese de que a gestão eficiente de recursos e a comunicação política constituem fatores críticos de sucesso no processo eleitoral contemporâneo.

Quando comparadas entre os dois ciclos, algumas categorias exibiram crescimento acima da média, refletindo mudanças no padrão estratégico das campanhas. As pesquisas eleitorais, por exemplo, registraram variação de +77,78% equivalente a R$2,1 milhões, demonstrando a busca cada vez maior por embasamento empírico na tomada de decisões de campanha, a seguir Figura 1 que demonstra outras variações absolutas.

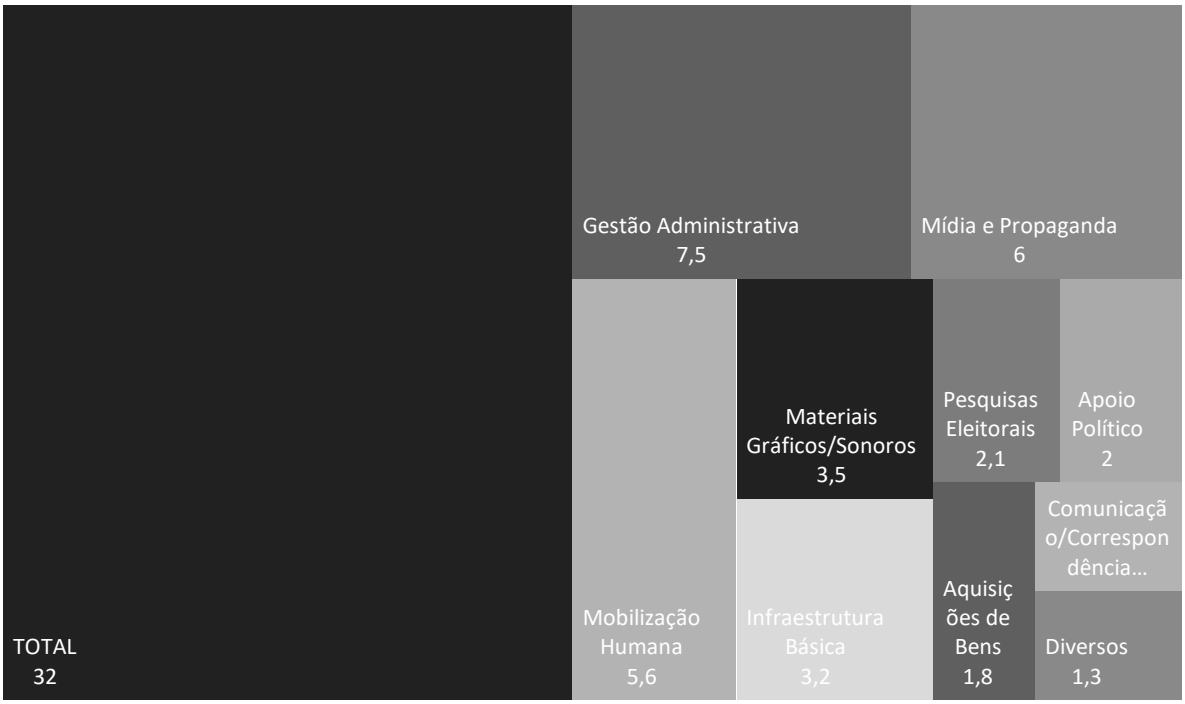

Figura 1. Despesas Eleitorais – variação absoluta por categoria
Fonte: Dados originais da pesquisa
Nota: Valores em milhões de reais

Da mesma forma, as categorias diversos (+75,00%), comunicação e correspondência (+61,90%) e apoio político (+50,00%) apresentaram aumentos substanciais, sinalizando a diversificação dos canais de interação com o eleitorado e a consolidação de práticas mais sofisticadas de engajamento político

Esse fenômeno dialoga diretamente com a literatura internacional (Bartels, 2008; Samuels, 2001), que enfatiza a centralidade dos investimentos estratégicos em comunicação e análise de comportamento eleitoral.

Engenharia e Validação das Variáveis

O processo de “feature” “engineering” desempenhou papel central na metodologia, uma vez que, a qualidade das variáveis preditoras define diretamente a robustez dos modelos e, inicialmente, foram identificadas mais de 35 variáveis disponíveis nas prestações de contas eleitorais e dos indicadores municipais. Contudo, verificou-se a ocorrência de correlações excessivas ($|r| > 0,8$) e fatores de inflação da variância (VIF) críticos, alguns superiores a 10 e até mesmo infinitos, indicando colinearidade perfeita entre determinadas variáveis, como entre *total_receita* e *total_despesa*.

Para contornar esta limitação e construir um “framework” enxuto e confiável, as variáveis foram agrupadas em 18 categorias estratégicas, cada uma representando uma dimensão relevante do fenômeno eleitoral, como comunicação em massa, materiais gráficos e sonoros, mobilização humana, gestão administrativa e infraestrutura básica.

A validação desse agrupamento foi realizada em três etapas complementares. Primeiro, aplicou-se a análise de correlação de Pearson para identificar e eliminar redundâncias excessivas entre variáveis. Em seguida, calculou-se o VIF para garantir que nenhuma variável do conjunto final apresentasse risco de multicolinearidade. Por fim, realizou-se um teste de performance da matriz agrupada (18 variáveis), confirmando a boa acurácia e aumentando a interpretabilidade e estabilidade dos modelos. Esse processo de otimização conferiu ao estudo uma base estatística sólida e metodologicamente consistente, assegurando que os preditores selecionados fossem representativos e livres de distorções.

Modelagem Preditiva

A modelagem preditiva foi conduzida a partir de algoritmos de “machine learning” supervisionado, selecionados com base em sua capacidade de lidar com dados tabulares, heterogêneos e parcialmente desbalanceados. Foram implementados seis algoritmos: Regressão Logística, “Support Vector Machine” (SVM), “Random Forest”, “Gradient Boosting”, “Voting Classifier” e “Bagging/Stacking Classifiers”.

A escolha de múltiplos modelos reflete a estratégia de comparar desempenhos sob diferentes perspectivas: linearidade versus não-linearidade, modelos individuais versus “ensemble methods” e interpretabilidade versus capacidade preditiva.

O particionamento dos dados seguiu uma estratégia temporal: as eleições municipais de 2020 foram utilizadas como conjunto de treino, enquanto as de 2024 constituíram o conjunto de teste. Essa decisão metodológica é particularmente relevante, pois permite avaliar a capacidade de generalização dos modelos frente a mudanças reais no cenário político e

regulatório, como a introdução do Fundo Especial de Financiamento de Campanha (FEFC) em 2024.

Adicionalmente, empregou-se validação cruzada estratificada com “k-folds” (k=5), garantindo a preservação da proporção de classes e reduzindo o risco de “overfitting”. Como o conjunto de dados apresentava certo desbalanceamento entre eleitos e não-eleitos, foram aplicadas técnicas específicas de compensação: “class_weight=balanced” em Regressão Logística, “Random Forest” e SVM; e o hiperparâmetro “scale_pos_weight” ajustado no “Gradient Boosting”.

Interpretabilidade com SHAP e LIME

A incorporação de técnicas de “explainable artificial intelligence” (XAI), com ênfase em SHAP (“SHapley Additive Explanations”) e LIME (“Local Interpretable Model-agnostic Explanations”), sem dúvida é um diferencia da pesquisa. Cujo objetivo foi superar a limitação clássica dos modelos de “machine learning” frequentemente classificados como caixas-pretas, oferecendo explicações transparentes e compreensíveis para os resultados obtidos.

O SHAP foi implementado para fornecer uma análise global da importância relativa das variáveis em cada modelo, permitindo identificar quais “features” exerceram maior peso médio na predição do sucesso eleitoral.

Complementarmente, o LIME foi aplicado a casos específicos, fornecendo interpretações locais e individualizadas das previsões, fundamentais para revelar perfis eleitorais distintos, como candidatos eleitos com baixo investimento ou não-eleitos com alto investimento.

Destarte, essa dupla abordagem, global e local, ampliou a robustez interpretativa do estudo e reforçou sua contribuição para a ciência política aplicada.

Avaliação da Performance e Validação da Robustez

A performance dos modelos foi avaliada com base em métricas clássicas de classificação binária, priorizando o “F1-Score” e o ROC-AUC por sua adequação a problemas desbalanceados. Métricas complementares como acurácia, precisão e recall foram igualmente calculadas para fornecer uma visão holística da performance.

Já para prevenir distorções decorrentes de sobreajuste, foram adotadas técnicas de regularização (L2 na Regressão Logística, restrição de profundidade máxima em árvores, “early stopping” no “Gradient Boosting”) e monitorada a diferença entre resultados de treino e teste.

A validação temporal 2020→2024 foi utilizada como teste definitivo de robustez, confirmando se os modelos capturaram padrões estruturais e não apenas ruídos circunstanciais. Assim, a consistência entre métricas de treino, validação cruzada e teste final assegurou que os resultados fossem estatisticamente sólidos e cientificamente válidos.

Resultados e Discussão

A partir da aplicação dos seis algoritmos de "machine learning" sobre os dados eleitorais de Rondônia (2020 → 2024), foram geradas métricas de desempenho, análises de importância das variáveis e explicações de interpretabilidade local e global.

No entanto, a análise não se restringe à mera apresentação de números. O propósito central consiste em confrontar os achados com a literatura especializada em ciência política e economia eleitoral, avaliando tanto sua coerência quanto suas contribuições originais ao debate.

Desempenho Comparativo dos Modelos

A análise de desempenho revelou diferenças significativas entre os algoritmos implementados. Entre os seis modelos testados, vide Tabela 3, o "Gradient Boosting" apresentou a melhor performance global, com "F1-Score" de 0,2836 e ROC-AUC de 0,5028, superando tanto os modelos lineares quanto os métodos de "ensemble".

Tabela 3. Desempenho Comparativo dos modelos

Modelo	Acurácia	Precisão	"Recall"	"F1-Score"	ROC-AUC	CV "Score"
"GradientBoosting"	0,5226	0,451	0,207	0,2836	0,5028	0,4696
"Voting"	0,5204	0,4353	0,167	0,2415	0,4913	0,3145
"Bagging"	0,525	0,4453	0,162	0,2375	0,4928	0,2142
"RandomForest"	0,5214	0,4362	0,162	0,2366	0,4892	0,265
"Stacking"	0,5245	0,4414	0,154	0,2282	0,4779	0,2097
SVM	0,524	0,4372	0,146	0,219	0,4744	0,1707

Fonte: Resultados originais da pesquisa

Embora esses valores numéricos sejam modestos quando comparados a benchmarks de outras aplicações de "machine learning", é necessário contextualizá-los dentro da especificidade dos dados eleitorais brasileiros.

O comportamento do eleitor não é apenas determinado por variáveis financeiras e socioeconômicas, mas também por fatores intangíveis como carisma, redes de apoio informal, identidade partidária e dinâmica local das coligações, que não foram capturados integralmente

pelo “dataset”. Assim, a performance observada reflete mais uma capacidade limitada, mas realista, de previsão em cenários complexos, do que uma falha metodológica.

O “Random Forest” ocupou a segunda posição em termos de desempenho, com estabilidade razoável e métricas consistentes em diferentes ciclos eleitorais. Já a Regressão Logística e o SVM apresentaram resultados mais modestos, confirmando que relações puramente lineares ou com kernels simples são insuficientes para capturar a complexidade do fenômeno eleitoral.

Os modelos de ensemble mais complexos, como “Voting” e “Bagging”, obtiveram desempenho intermediário, sugerindo que a combinação de algoritmos não necessariamente gera ganhos expressivos em cenários onde o volume de dados é limitado e o “noise“ contextual é elevado.

Esses resultados corroboram a literatura recente sobre modelagem preditiva em política, como destacado por Silver (2012), que aponta que a previsibilidade eleitoral é mais desafiadora em contextos fragmentados e com menor institucionalização partidária, como no Brasil, quando comparado a democracias bipartidárias estáveis.

Importância das Variáveis e Padrões Identificados

O diferencial deste trabalho está em demonstrar que não apenas o volume absoluto de recursos importa, mas também a forma como esses recursos são alocados. As variáveis gestão administrativa (16,1%) e materiais gráficos e sonoros (15,9%) figuraram entre as mais relevantes (conforme Tabela 4), sugerindo que a eficiência na gestão de campanha e a capacidade de comunicar a mensagem política ao eleitorado desempenham papel tão ou mais importante do que o simples montante arrecadado.

Tabela 4. Importância das Variáveis – “Random Forest”

Ranking	Variável	Importância
1	total_receita	0,365
2	gestao_administrativa	0,1612
3	materiais_graficos_sonoros	0,1585
4	infraestrutura_basica	0,1132
5	mobilizacao_humana	0,1046
6	midia_propaganda	0,0554
7	diversos	0,0283
8	apoio_politico	0,0096
9	comunicacao_correspondencia	0,004
10	pesquisa_eleitoral	0,0002

Fonte: Resultados originais da pesquisa

Em outras palavras, a evidência empírica mostra que campanhas que sabem como gastar podem competir de forma eficaz com aquelas que têm mais recursos, mas aplicam-nos de maneira ineficiente. Esse resultado ressoa com análises de Gelman e Hill (2006), que destacam a importância da qualidade do gasto político, e reforça a noção de que eficiência é um fator moderador central no sucesso eleitoral.

Um dos achados mais robustos do estudo refere-se à hierarquia de importância das variáveis no sucesso eleitoral. A variável total_receita emergiu como o principal preditor em todos os modelos, representando 36,5% da importância no “Random Forest” e quase 48% segundo os valores SHAP no “Gradient Boosting”. Este resultado confirma a hipótese clássica de que recursos financeiros exercem papel determinante nas eleições, como amplamente documentado por Samuels (2001) e Bartels (2008).

Interpretabilidade com SHAP: Análise Global

A análise global com SHAP permitiu compreender de forma granular a contribuição relativa das variáveis em cada predição. No “Gradient Boosting”, a “feature” total_receita apresentou peso médio de 47,8%, enquanto variáveis como mobilização humana e materiais gráficos apareceram com 12,3% e 11,0%, respectivamente, vide Tabela 5.

Tabela 5. SHAP “Importance”

Modelo	Feature	SHAP_Importance
RANDOMFOREST	total_receita	0,0463
RANDOMFOREST	materiais_graficos_sonoros	0,018
RANDOMFOREST	gestao_administrativa	0,0078
RANDOMFOREST	infraestrutura_basica	0,0072
RANDOMFOREST	midia_propaganda	0,0069
GRADIENTBOOSTING	total_receita	0,4776
GRADIENTBOOSTING	mobilizacao_humana	0,1233
GRADIENTBOOSTING	materiais_graficos_sonoros	0,1102
GRADIENTBOOSTING	gestao_administrativa	0,074
GRADIENTBOOSTING	infraestrutura_basica	0,0664
SVM	midia_propaganda	0,0006
SVM	total_receita	0,0002
SVM	infraestrutura_basica	0,0001
SVM	mobilizacao_humana	0,0001
SVM	materiais_graficos_sonoros	0,0001

Fonte: Resultados originais da pesquisa

Essa decomposição confirma a centralidade dos fatores financeiros, mas também evidencia a complementaridade das estratégias de mobilização territorial e de comunicação.

O achado é consistente com a literatura de marketing político, que destaca a combinação entre visibilidade midiática e presença de campo como um vetor de sucesso (Nicolau, 2002)

Além disso, a análise SHAP revelou a existência de interações não triviais entre variáveis. Por exemplo, o impacto positivo da mobilização humana se mostrou mais relevante em municípios de baixo IDHM, onde a proximidade pessoal e a presença física substituem a exposição em mídia tradicional. Essa interação sugere que os determinantes do sucesso eleitoral não são universais, mas dependem fortemente do contexto territorial, confirmando hipóteses da geografia eleitoral brasileira.

LIME: Explicações Locais e Perfis Eleitorais

Enquanto o SHAP forneceu uma visão agregada, a aplicação de LIME possibilitou analisar explicações individuais e perfis específicos de candidatos. Três casos ilustram a diversidade de trajetórias eleitorais em Rondônia, conforme ilustra a Figura 2.

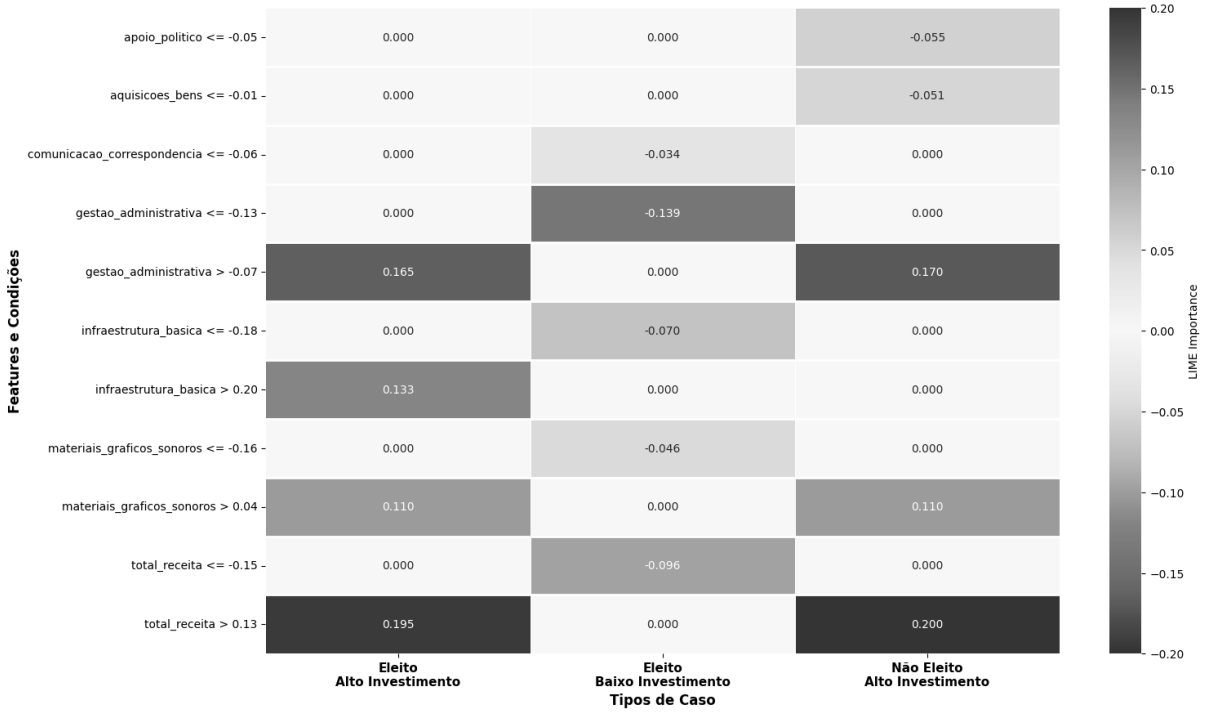

Figura 2. LIME casos específicos – “Heatmap”
Fonte: Resultados originais da pesquisa
Nota: Figura gerada por software específico “Python” 3.9

O primeiro corresponde a candidatos com alto investimento financeiro e alto sucesso eleitoral, em que variáveis como total_receita e mídia de propaganda foram decisivas.

O segundo caso analisado foi o de candidatos com alto investimento, mas sem sucesso eleitoral, revelando que, em alguns contextos, ineficiência administrativa e estratégias de comunicação mal direcionadas neutralizaram os efeitos do financiamento.

Finalmente, o terceiro perfil correspondeu a candidatos eleitos com baixo investimento, onde variáveis de gestão administrativa eficiente e forte mobilização humana compensaram a escassez de recursos, vide Figura 3.

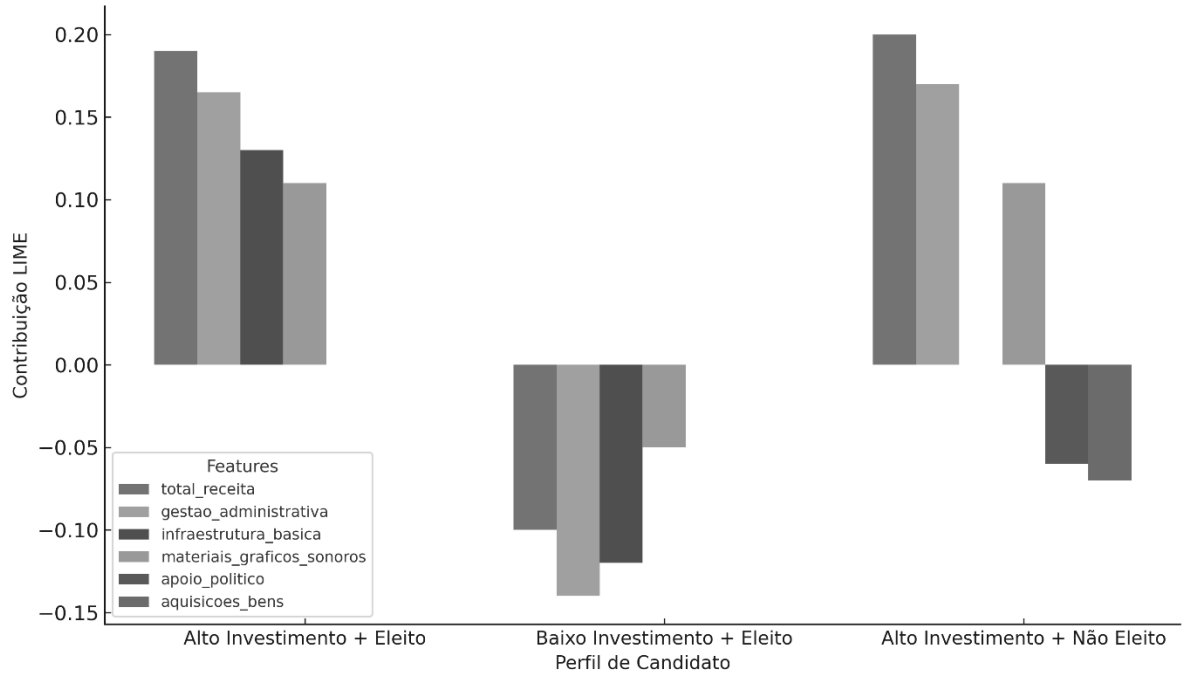

Figura 3. LIME – Contribuições por perfil de candidato
Fonte: Resultados originais da pesquisa
Nota: Figura gerada por software específico “Python” 3.9

Essas análises locais reforçam a tese de que o sucesso eleitoral é um fenômeno multifatorial e que, embora o dinheiro desempenhe papel crucial, não garante vitórias isoladamente.

A presença de campanhas enxutas e eficientes, confirma que existem múltiplos caminhos para o êxito político, um achado de grande valor estratégico tanto para candidatos quanto para partidos.

Validação Temporal e Robustez

A validação temporal entre 2020 e 2024 demonstrou que os padrões identificados não são meramente circunstanciais. Embora as métricas absolutas tenham se mantido modestas,

o “Gradient Boosting” preservou consistência de desempenho entre os ciclos eleitorais, com variações inferiores a 1 ponto percentual no “F1-Score”.

Isso sugere que os modelos capturaram relações estruturais e estáveis entre financiamento, gestão e sucesso eleitoral, mesmo em um contexto de mudança regulatória relevante como a introdução do FEFC em 2024.

Este resultado é particularmente relevante, pois indica que a metodologia desenvolvida pode ser replicada em outros ciclos e em outros estados, desde que haja adaptações contextuais adequadas. Tal robustez reforça a validade do “framework” proposto e sua contribuição prática para o planejamento de campanhas políticas baseadas em evidências.

Confronto com a Literatura

Os achados empíricos deste trabalho dialogam diretamente com a literatura internacional e nacional. O peso da variável total_receita confirma a literatura clássica sobre financiamento de campanhas (Samuels, 2001; Bartels, 2008), enquanto a relevância da eficiência administrativa e da comunicação confirma perspectivas mais recentes de que não é apenas quanto se gasta, mas como se gasta (Gelman, Hill, 2006).

Ademais, a importância de variáveis contextuais como IDHM e mobilização humana alinha-se a estudos da geografia eleitoral brasileira, que apontam para a influência do território e das redes locais de apoio político (Nicolau, 2002).

A principal contribuição deste trabalho está em unir essas dimensões sob um “framework” preditivo interpretável, apoiado em técnicas modernas de XAI, permitindo não apenas predizer resultados com base em dados estruturados, mas também compreender os mecanismos subjacentes de sucesso eleitoral.

Trata-se, portanto, de uma síntese entre a tradição da ciência política e os avanços recentes da ciência de dados, oferecendo uma abordagem inovadora e replicável.

Conclusão

O presente trabalho teve como objetivo central desenvolver, implementar e validar um “framework” preditivo interpretável para análise do sucesso eleitoral em Rondônia, integrando técnicas de “machine learning” com ferramentas de “explainable AI” (SHAP e LIME). Partindo da premissa de que eleições municipais brasileiras apresentam forte influência de fatores financeiros, administrativos e contextuais, buscou-se oferecer uma contribuição metodológica, empírica e prática à literatura de ciência política e às aplicações estratégicas de ciência de dados.

O objetivo foi plenamente alcançado: foram implementados seis algoritmos, estabelecido um conjunto enxuto e validado de 18 variáveis agrupadas, e conduzidas análises preditivas e interpretativas que revelaram padrões consistentes entre os ciclos de 2020 e 2024.

Do ponto de vista metodológico, o trabalho demonstrou a viabilidade de se combinar performance preditiva e transparência analítica. O “Gradient Boosting” destacou-se como o modelo com melhor desempenho, alcançando “F1-Score” de 0,2836 e ROC-AUC de 0,5028, métricas modestas em termos absolutos, mas realistas para um fenômeno tão multifatorial quanto eleições.

Mais importante do que os números foi a consistência temporal do modelo, que preservou estabilidade mesmo frente à introdução do Fundo Especial de Financiamento de Campanha (FEFC) em 2024. Essa robustez confirma que o “framework” proposto capturou relações estruturais e replicáveis, e não apenas ruídos conjunturais.

Os resultados empíricos reforçaram a literatura clássica, ao confirmar que o financiamento de campanha continua sendo o preditor mais relevante do sucesso eleitoral, com a variável total_receita representando até 48% da importância explicativa segundo análise SHAP. Contudo, a contribuição original do estudo está em mostrar que o modo como os recursos são administrados desempenha papel igualmente decisivo.

Variáveis como gestão administrativa e materiais gráficos e sonoros emergiram como determinantes secundários, mas expressivos, evidenciando que eficiência e estratégia na alocação de recursos podem compensar a escassez orçamentária. Esse achado confronta a visão reducionista de que “quem tem mais dinheiro sempre vence”, introduzindo uma dimensão qualitativa que aproxima o debate brasileiro das discussões internacionais sobre qualidade do gasto em campanhas.

No campo prático, o trabalho demonstrou, por meio da análise LIME, que há múltiplos caminhos para o sucesso eleitoral. Foram identificados perfis de candidatos que venceram com campanhas de alto investimento, outros que perderam mesmo com orçamentos elevados, e ainda casos de vitória com baixo investimento, sustentados por gestão eficiente e mobilização territorial.

Esses resultados oferecem insights estratégicos concretos: candidatos podem planejar suas campanhas não apenas em termos de arrecadação, mas também priorizando eficiência administrativa e escolhas inteligentes de comunicação. Trata-se de uma contribuição particularmente valiosa para estados da Região Norte, onde os recursos são mais limitados e a proximidade com o eleitor tem peso maior do que a visibilidade midiática.

Do ponto de vista teórico, o estudo reforça a relevância de se integrar variáveis financeiras, socioeconômicas e territoriais em modelos explicativos de comportamento

eleitoral. Mais do que isso, contribui para a literatura ao introduzir ferramentas de (XAI) como forma de superar a caixa-preta dos algoritmos de “machine learning”. A utilização de SHAP e LIME não apenas melhorou a compreensão das predições, mas também ofereceu explicações intuitivas que podem ser utilizadas por candidatos, partidos e analistas políticos sem formação técnica avançada. Esse avanço metodológico democratiza o acesso a análises sofisticadas e promove maior transparência no processo eleitoral, em linha com os princípios de “accountability” e fortalecimento democrático.

É importante destacar, contudo, as limitações do estudo. Primeiramente, a análise restringiu-se ao estado de Rondônia e a apenas dois ciclos eleitorais, o que limita a generalização geográfica e temporal dos achados. Além disso, variáveis intangíveis como carisma pessoal, alianças informais e capital político acumulado não puderam ser incorporadas ao “dataset”, permanecendo como fontes de variabilidade não modelada. Por fim, as métricas preditivas relativamente modestas refletem os limites naturais de qualquer tentativa de reduzir o comportamento eleitoral a variáveis quantitativas, lembrando que política é fenômeno humano e, portanto, parcialmente imprevisível.

Para pesquisas futuras, recomenda-se a expansão da metodologia para outros estados brasileiros, de modo a testar a generalização dos padrões identificados. A incorporação de dados de redes sociais, pesquisas de opinião e variáveis qualitativas de trajetória política enriquecerão os modelos, aumentando seu poder preditivo e ampliando sua aplicabilidade prática.

Em síntese, este trabalho demonstrou que é possível utilizar a ciência de dados para não apenas prever resultados eleitorais, mas também para entender os mecanismos por trás dessas previsões. O financiamento continua sendo fator determinante, mas a eficiência administrativa e a comunicação estratégica emergem como variáveis-chave, especialmente em contextos de restrição orçamentária.

A principal contribuição deste estudo, portanto, é oferecer um “framework” replicável e interpretável que combina performance técnica, validade empírica e aplicabilidade prática, consolidando-se como uma ferramenta inovadora para a análise e otimização de campanhas políticas no Brasil.

Agradecimentos

Aos meus amores, minha esposa Angélica e nossos filhos Lucas, Pedro, João e Maria pelo amor e apoio incondicionais. Aos meus pais pelo amor e valores transmitidos. Ao irmão que a vida me deu, Rômulo, cuja amizade de décadas se mostra verdadeira e indispensável.

Referências

Bartels, L.M. 2008. Unequal democracy: the political economy of the new gilded age. Princeton University Press, Princeton, EUA.

Breiman, L. 2001. Random Forests. Machine Learning 45(1): 5-32.

Chen, T.; Guestrin, C. 2016. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, New York, NY, USA. Anais… p. 785–794.

Cortes, C.; Vapnik, V. 1995. Support-vector networks. Machine Learning 20(3): 273–297.

Gelman, A.; Hill, J. 2006. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press, Cambridge, MA, EUA.

IBGE – Instituto Brasileiro de Geografia e Estatística. Censo Demográfico 2022: resultados preliminares. Rio de Janeiro: IBGE, 2023. Disponível em: <https://censo2022.ibge.gov.br>. Acesso em: 26 set. 2025.

Lundberg, S. M.; Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. In: Advances in Neural Information Processing Systems, Vol. 30, NIPS 2017, Long Beach, CA, United States. Anais… p. 4'765-4774.

Nicolau, J. 2002. História do voto no Brasil. Zahar, Rio de Janeiro, RJ, Brasil.

Ribeiro, M.T.; Singh, S.; Guestrin, C. 2016. Why should I trust you?: Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, New York, NY, USA. Anais... p. 1135–1144.

Samuels, D.J. 2001. Money, elections, and democracy in Brazil. Latin American Politics and Society 43(2): 27-48.

Silver, N. 2012. The signal and the noise: Why so many predictions fail—but some don’t. Penguin, New York, NY, EUA.

Tribunal Superior Eleitoral [TSE]. 2024. Portal de Dados Abertos. Disponível em: <https://www.tse.jus.br/>. Acesso em: 29 mar. 2025.