Proposta de um modelo para predição do resultado das eleições presidenciais brasileiras baseado em técnicas de regressão

1st Gabriel P. Robaina, 2nd Fabiano Baldo
Departamento de Ciência da Computação - DCC
Programa de Pós-Graduação em Computação Aplicada - PPGCAP
Universidade do Estado de Santa Catarina - UDESC
Joinville, SC, Brasil
gabriel.robaina@edu.udesc.br, fabiano.baldo@udesc.br

***Abstract*—Prediction of elections is a subject that excites the population, especially in the last few months before an election. In Brazil, there is a wide availability of political, economic and social data, in institutions such as TSE, IBGE and opinion research institutes that can be used as sources to create prediction models. Therefore, this work aims to build multivariate linear regression and regression tree models to predict the percentage of votes received by the situational candidate for the presidency of Brazil. The multivariate linear regression model had the smallest prediction errors, with MAE of 1.45 in the first round and 1.48 in the second, with margins smaller than 1% in 2002, 2006 and 2018. The proposed models seemed to be more accurate than other models found in the literature. As main contributions, it was possible to observe that the sampling of data by state and the use of the illiteracy rate and the popular vote intention contributed directly to the performance of the models.**

I. Introdução

A ditadura militar brasileira teve início em 1945 e perdurou até a eleição do civil Tancredo Neves como Presidente da República pelo Colégio Eleitoral em 1985 [Dias 2016]. No entanto, o processo de redemocratização do país foi finalizado posteriormente com uma nova constituição que estabeleceu a sistemática em vigor para as eleições presidenciais [Brasil 1988], [Brasil 1965]. As eleições presidenciais são eleições majoritárias em dois turnos, onde no primeiro todos os candidatos disputam o pleito e, caso nenhum tenha obtido a maioria absoluta dos votos, é realizado um segundo turno de votação com os dois candidatos mais votados do primeiro.

Ainda, o Código Eleitoral delega a Justiça Eleitoral brasileira as competências de organizar o eleitorado nacional e gerir o processo eleitoral. Fazem parte da Justiça Eleitoral o Tribunal Superior Eleitoral (TSE), com jurisdição em todo o país, um Tribunal Regional Eleitoral (TRE), por estado, e as juntas e juízes eleitorais [Brasil 1965].

Nesse contexto, a lei nº 12.527/2011 regulamenta que é de responsabilidade dos orgãos públicos propiciar o amplo acesso e divulgação da informação [Brasil 2011]. Portanto, o TSE disponibiliza uma base de dados eleitorais do ano de 1945 em diante, onde constam receitas e despesas de campanha dos candidatos e partidos, o perfil do eleitorado e a quantidade de votos que cada candidato recebeu em um município, zona eleitoral, ou mesmo urna eletrônica [TSE 2021].

Outras entidades também contribuem com acesso a dados de interesse público no Brasil. O Instituto Brasileiro de Geografia e Estatística (IBGE) disponibiliza dados econômicos, como o crescimento do produto interno bruto (PIB) e da inflação [IBGE 2021b], [IBGE 2021d], e sociais, como a taxa de analfabetismo nas regiões brasileiras [IBGE 2021c], [IBGE 2021a]. Ainda, institutos de pesquisa, como o Datafolha, realizam e divulgam periodicamente resultados de pesquisas de intenção de voto das eleições [Datafolha 2021].

No contexto do aprendizado de máquina, o aprendizado supervisionado é uma técnica que permite a construção de modelos preditivos a partir de uma base de dados de treinamento. Os dados de treinamento são compostos de entradas conhecidas e saídas esperadas que são apresentados para o método de indução que constrói o modelo que generaliza o conhecimento contido na base de treinamento. Após o treinamento do modelo, testes são realizados para medir a capacidade de representação do conhecimento contida nele, e que será utilizado para fazer a predição de novas instâncias [Marsland 2009].

A regressão é uma técnica de aprendizado de máquina supervisionado que permite a predição de uma classe numérica contínua por meio da combinação dos atributos de entrada [Witten et al. 2007]. No contexto das eleições presidenciais brasileiras, os atributos de entrada podem ser representados por dados políticos, econômicos e sociais, cujo o objetivo é prever a porcentagem de votos recebidos por um candidato. Diante da ampla disponibilidade de dados sobre fatores que impactam diretamente no resultados de eleições majoritárias, este trabalho apresenta a seguinte pergunta de pesquisa: Como prever adequadamente o resultado das eleições presidenciais brasileiras por meio da utilização de modelos de regressão?

A. Objetivo geral

Construir um modelo de regressão capaz de prever resultados de eleições presidenciais brasileiras com acurácia

satisfatória.

B. Objetivos específicos

1) Construir uma base de dados a partir das informações disponibilizadas pelo TSE, pelo IBGE e pelos institutos de pesquisa.
2) Treinar modelos de regressão e selecionar o mais preciso na previsão do resultado das eleições presidenciais.

C. Metodologia

A presente metodologia se baseia nos nove passos do processo clássico de descoberta de conhecimento em bases de dados [Fayyad et al. 1996]. Primeiro, as definições do processo eleitoral brasileiro foram estudadas para compreensão das informações presentes nas bases do TSE e dos institutos de pesquisa. Depois, um recorte temporal dos dados das eleições brasileiras foi definido com base na abrangência, completude e disponibilidade deles sobre o processo eleitoral. Os dados do recorte foram coletados, reunidos e pré-processados em uma base para a remoção de duplicatas, valores ausentes e anomalias. A regressão foi escolhida como método de aprendizagem de máquina para a tarefa de predição do número de votos de um candidato, e os dados resultantes do recorte definem o conjunto base de treinamento e validação dos modelos.

As técnicas de regressão linear multivariada e árvore de regressão foram escolhidas para a tarefa de regressão, pois a variável objetivo da mineração é o número de votos obtidos pelos candidatos, portanto, uma variável contínua. Dados da eleição de 2018 não fizeram parte do conjunto de treinamento dos modelos, pois foram reservados para o teste de predição fora da amostra. Portanto, os modelos foram treinados com dados das eleições presidenciais anteriores a 2018 que estão contidas no recorte. Ainda, os modelos foram testados e avaliados dentro do conjunto de treinamento com o método de validação cruzada *10-fold*. Depois, os modelos construídos foram aplicados no teste de predição, tendo sua performance avaliada com os dados de 2018, não utilizados no treinamento.

A performance dos modelos foi quantificada durante a etapa de validação cruzada utilizando o coeficiente de determinação (R^2) e o erro médio absoluto (MAE - *mean absolute error*). A precisão dos modelo foi avaliada no teste de predição pelo MAE na comparação com votações reais das eleições presidenciais. Essa fase envolveu correções nas etapas anteriores visando o refinamento dos resultados. Por fim, os resultados, limitações da abordagem e contribuições foram discutidas e documentadas.

II. Revisão de conceitos e trabalhos relacionados

A. Regressão linear multivariada

A regressão linear multivariada modela o relacionamento entre uma variável dependente *Y* e múltiplas variáveis independentes *X* por meio de um hiperplano, onde *a* e *b* representam, respectivamente, sua constante (intercepto) e inclinações [Russell and Norvig 2004], conforme equação geral apresentada em (1) considerando k variáveis independentes:

$$Y = a_0 + b_1X_1 + b_2X_2 + ... + b_kX_k \tag{1}$$

Dentro de um conjunto de dados, é possível encontrar o hiperplano que melhor representa o relacionamento linear entre as variáveis por meio do método dos mínimos quadrados, minimizando a soma dos erros de predição ao quadrado (SSE - *sum of squared estimate of errors*) [Lewis-Beck 2015]. Ainda, a equação geral da regressão linear multivariada pode ser representada em forma matricial, conforme (2), sendo e o erro estatístico associado. Nessa representação, considerando n observações de dados de treinamento, Y e e são vetores $n \times 1$, X é uma matriz $n \times (k+1)$ e b um vetor $(k+1) \times 1$ [Weisberg 2005].

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix} * \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} \tag{2}$$

A estimativa de constantes por mínimos quadrados ordinários (OLS - *ordinary least squares*) estabelece que o vetor de coeficientes b que minimiza o SSE pode ser descrito por (3) [Weisberg 2005].

$$\mathbf{b} = (\mathbf{X^T X})^{-1}\mathbf{X^T Y} \tag{3}$$

B. Árvore de regressão

As árvores de regressão são baseadas na estratégia de dividir para conquistar e são aplicadas quando a variável objetivo da predição é contínua. Cada nó da árvore divide o conjunto de dados em dois subconjuntos com base nos valores de um determinado atributo [Witten et al. 2007]. Essa divisão tem a finalidade de minimizar a variância da variável dependente nos subconjuntos filhos. A medida de variância na divisão de um conjunto D em dois subconjuntos D_L e D_R, respectivamente à esquerda e à direita do nó original, pode ser calculada pela redução do desvio padrão SDR (4), onde sd é o desvio padrão dos conjuntos n, n_L e n_R que representam, respectivamente, os tamanhos do conjunto original e dos subconjuntos à esquerda e à direita [Faceli et al. 2011]. A divisão em subconjuntos prossegue até que a redução no desvio padrão em uma divisão seja menor que um limite parametrizado no treinamento [Witten et al. 2007].

$$SDR = sd(D, y) - \frac{n_L}{n} \times sd(D_{L,y}) - \frac{n_R}{n} \times sd(D_{R,y}) \tag{4}$$

Na árvore de regressão as folhas são representadas pela média da variável dependente nos subconjuntos, e seus valores atribuídos às instâncias submetidas ao modelo [Witten et al. 2007].

A etapa de poda acontece após a construção do modelo e visa diminuir o superajustamento convertendo subárvores em folhas. A poda por redução de erro compara o erro estático com o erro de *backed-up* nos nós. O erro estático de um nó pode ser medido pelo erro de predição da variável dependente

de um exemplo caso seja convertido em folha. O erro de *backed-up* de um nó, por sua vez, está associado ao erro de predição realizado pela sua subárvore. Se o erro estático for menor ou igual ao erro de *backed-up*, então não existe ganho de performance do modelo associado à existência da subárvore e, portanto, ela pode ser substituída por uma folha [Faceli et al. 2011].

C. Métricas de avaliação da predição

O MAE e o coeficiente de determinação (R^2) foram selecionados para a avaliação das predições. MAE representa a média dos módulos dos erros individuais de predição sem considerar os sinais. Já o coeficiente de determinação R^2 é calculado a partir do coeficiente de correlação R, variando entre 0 e 1, e pode ser descrito como a porcentagem da variação dos valores reais que está sendo explicada pelo preditor [Witten et al. 2007], [Lewis-Beck 2015].

O MAE foi selecionado como métrica de avaliação por fornecer uma medida facilmente interpretável, representando, por exemplo, a média de erro do percentual de voto recebido por um candidato da situação em um determinado estado. O coeficiente de determinação R^2, por outro lado, indica como o preditor explica a variação dos votos recebidos pelo candidato nos estados e é uma métrica de regressão que se aproxima, em semântica, da precisão em uma tarefa de classificação. O uso do R^2 também possibilita a comparação dos experimentos com outros trabalhos relacionados que utilizaram essa mesma métrica.

III. Trabalhos relacionados

Abordagens estatísticas do problema de predição de eleições vem sendo empregadas principalmente em democracias consolidadas como os Estados Unidos, Reino Unido e França, e estabelecem o voto no candidato da situação como dependente da sua popularidade, que é um reflexo do seu desempenho político, e do cenário econômico [Lewis-Beck 2005]. Classicamente essas abordagens envolvem modelos de regressão construídos a partir de instâncias de eleições nacionais com dados macroeconômicos e de pesquisas de intenção ou expectativa de voto de um período anterior a eleição [Magalhães et al. 2012], [Lewis-Beck 2005].

Os modelos construídos para essa finalidade devem equilibrar precisão e *lead* [Jennings et al. 2020]. A precisão mede a capacidade que o modelo tem de prever o resultado de uma eleição, enquanto o *lead* mede o tempo de antecedência em relação ao pleito com que essa predição pôde ser realizada. Por exemplo, um modelo baseado em pesquisas eleitorais de véspera de eleição tende a ter precisão, mas não tem *lead*, já que uma predição feita perto do dia da eleição é trivial. Em geral, uma diminuição de *lead* gera um aumento de precisão. Um tempo de *lead* de três meses a um ano é considerado satisfatório e não gera perdas significativas de precisão [Jennings et al. 2020].

A pesquisa eleitoral é uma das ferramentas mais utilizadas na predição de eleições [Jennings et al. 2020]. Lozano e Castillo [Lozano and Castillo 2008] utilizaram dados socioeconômicos, políticos e de intenção de voto na Espanha para construir um modelo de árvore de decisão que classifica o resultado das eleições com mais de 90% de precisão, e se apresenta como ferramenta complementar à pesquisa. Ainda, sob o ponto de vista dos autores, a análise do modelo demonstra que a memória do voto de um indivíduo em eleições passadas e sua ideologia política são determinantes na escolha do candidato.

Na França, Nadeau et al. [Nadeau et al. 2010] modelaram o voto como dependente apenas da popularidade, que por sua vez depende de variáveis políticas e econômicas. Dessa forma, dois modelos de regressão, construídos utilizando mínimos quadrados ordinários (OLS - *orginary least squares*), foram utilizados: O primeiro, para predição de votos na situação, atingiu um MAE de 2.3 no conjunto de testes a partir do índice de popularidade nas pesquisas eleitorais com *lead* de 6 meses. O segundo comprovou, com um R^2 de 0.73, que os índices de desemprego, tempo da situação no poder e a diferença entre a aprovação do presidente e primeiro ministro durante três cohabitações tem influência significativa na popularidade. Essa abordagem de dois modelos se mostrou um atenuante do número reduzido de 8 amostras de eleições presidenciais e do dinamismo do cenário político francês.

Na Espanha, o número baixo de observações de eleições nacionais também se mostrou um problema para a construção de um modelo de predição acurado das eleições do legislativo e do parlamento europeu [Magalhães et al. 2012]. As variáveis econômicas, como a inflação e o desemprego, e variáveis políticas, como a avaliação popular do governo no poder, foram utilizadas na construção de modelos de regressão com constantes determinadas por meio do método OLS. O modelo mais acurado foi aquele que levou em conta a relação das variáveis com o alinhamento ideológico do partido no poder: Se o partido de esquerda estivesse no poder, o desemprego não teria influência na porção de votos recebida pelo candidato da situação. Dessa forma, o modelo atingiu um R^2 de 0.95 nas eleições anteriores a 2011. Nesse ano, o modelo teve um baixo desempenho já que foi uma eleição atípica, com o pior resultado histórico do partido da situação até aquele momento.

A abordagem baseada em dados também pode ser utilizada na criação de modelos para predição de eleições globalmente. Kennedy et al. [Kennedy et al. 2017] utilizaram dados econômicos e políticos de 86 países para construir um modelo de árvore de regressão aditiva Bayesiana capaz de prever eleições de nível nacional com cerca de 80% de precisão em múltiplos países. A aplicação do modelo indicou um padrão de vantagem do candidato da situação em regimes políticos fechados e a importância da política externa na determinação do vencedor, aqui representada pela proximidade com os Estados Unidos e o volume de ajuda financeira estrangeira recebida. Dentre os dados utilizados destacam-se o PIB, crescimento econômico, grau de democracia e a predição feita pela pesquisa popular popular.

Turgeon e Rennó [Turgeon and Rennó 2012] apresentaram a escassez de dados disponíveis como um problema para as predições de eleições nacionais no Brasil e em democracias

jovens. Para mitigar essa limitação, utilizou-se a abordagem de regressão para predição do resultado de eleições presidenciais, com dados em nível estadual de crescimento do PIB e popularidade do candidato da situação entre 1994 e 2006. Foram aplicados modelos de regressão linear multivariada baseados em eleições francesas e americanas atingindo, no melhor preditor, um R^2 de 0,61 com SEE de 10,85 e com *lead* de 2 meses. O pior desempenho dos modelos foi observado na predição das eleições de 2010, já que havia sido a primeira vez que uma candidata da situação venceu após dois mandatos consecutivos do mesmo partido. Ainda, a escassez de dados de popularidade dos candidatos no nível estadual e o atraso da disponibilização de informações de crescimento do PIB nos estados foram apresentados como problemas para a construção de um modelo acurado.

Em suma, a regressão se mostra como uma técnica amplamente utilizada na tarefa de predição de eleições, representando a quantidade de votos no candidato da situação como variável dependente. Ainda, essas abordagens selecionam dados políticos, econômicos e de popularidade como variáveis independentes do modelo, e utilizam observações de eleições nacionais para determinação de suas constantes, limitando assim o número de instâncias de treinamento. Além disso, essa abordagem pode apresentar problemas relacionados à escassez de disponibilidade de dados principalmente em democracias jovens e emergentes. Por outro lado, a utilização de observações no nível estadual também apresentou dificuldades relacionadas a disponibilidade de dados nessa granularidade.

IV. Desenvolvimento

O recorte de dados selecionado priorizou o período pós redemocratização, e compreende o primeiro e segundo turnos das eleições presidenciais brasileiras para presidente realizadas no período de 2002 à 2018. O pleito de 1989 foi retirado do recorte por seu caráter excepcional [Turgeon and Rennó 2012]. As eleições de 1994 e 1998, por outro lado, foram excluídas do recorte pela escassez de dados relacionados ao financiamento de campanha que estão disponíveis apenas no período posterior a 2002. A eleição presidencial do ano de 2018 foi selecionada para o teste de predição e, portanto, não foi utilizada no treinamento dos modelos.

Dados econômicos, políticos, sociais e de intenção de voto foram selecionados e utilizados na construção da base de dados. Os atributos, assim como suas descrições e fontes, foram apresentados na Tabela I. As variáveis *dummy* representam uma categoria e podem receber exclusivamente os valores 0 e 1. Em um exemplo, o *dummy* de reeleição recebe 1 quando o candidato da situação à presidência está concorrendo ao seu segundo mandato. Ainda, a variável que representa as unidades federativas varia entre 1 e 27. A variável dependente selecionada foi a porcentagem de votos recebidos pelo candidato da situação por estado. Dessa forma as previsões dos modelos em cada estado podem ser consolidadas em uma votação nacional. Ainda, os dados do primeiro e segundo turnos foram separados em bases diferentes, possibilitando assim a construção de modelos específicos para cada turno.

Atributo	Período	Nível	Fonte
Despesa máxima de campanha	Eleição corrente	Candidato(a)	TSE
Crescimento percentual do PIB	Relativo ao ano anterior	Por estado	Consolidação do PIB dos municípios (IBGE)
Taxa de analfabetismo	Eleição corrente	Por região	SIS e PNAD (IBGE)
Crescimento percentual da inflação	Relativo à eleição anterior	Nacional	Série histórica IPCA (IBGE)
Porcentagem de intenções de voto	Eleição corrente, com *lead* variável para cada eleição e turno	Nacional	Pesquisas de intenção de voto (Datafolha)
Dummy reeleição	Eleição corrente	Candidato(a)	Conhecimento do domínio
Dummy terceiro mandato consecutivo do partido	Eleição corrente	Candidato(a)	Conhecimento do domínio
Unidade da Federação	Eleição corrente	Por estado	Conhecimento do domínio

TABLE I
DETALHAMENTO DOS ATRIBUTOS SELECIONADOS PARA O RECORTE DE DADOS.

Dados de taxa de analfabetismo para o ano de 2010 não estavam disponíveis no SIS e PNAD do IBGE [IBGE 2021c], [IBGE 2021a], portanto, a taxa considerada para esse ano foi a média entre os anos de 2009 e 2011. Ainda, não haviam dados de despesa máxima de campanha na base do TSE para a candidata da situação à presidência em 2014, portanto, consideramos para cada turno a metade da despesa total declarada pelo partido.

Os dados foram normalizados em uma escala de 0 a 1 durante a etapa de pré-processamento. Todos os atributos selecionados são numéricos, eliminando assim a necessidade de qualquer procedimento de conversão de classes ou símbolos.

As técnicas de regressão linear multivariada e árvore de regressão foram escolhidas para a tarefa de predição do percentual de votos recebidos pelo candidato da situação em cada estado. O Weka 3.8.5 [Frank et al. 2016] foi utilizado como plataforma de mineração de dados. Os preditores LinearRegression e REPTree foram selecionados na plataforma para as tarefas de regressão linear multivariada e árvore de regressão, respectivamente.

Todas as eleições do recorte envolveram votações em segundo turno. Por isso, foram construídos modelos com especialização em um turno da eleição, totalizando o treinamento de dois modelos de cada preditor, sendo um especializado no primeiro turno e o outro no segundo.

A consolidação das predições de nível estadual em nacional possibilita o cálculo da precisão dos modelos em relação às votações reais para presidente. Esse processo começa com a multiplicação do percentual de votos previstos pelo total de votos do estado. Depois, soma-se o número absoluto de votos em todos os estados, e calcula-se o percentual que ele representa da votação total em cada ano.

As métricas de avaliação dos modelos foram computadas

pela validação cruzada *10-fold* dentro do conjunto de treinamento, com dados de 2002 a 2014. Depois, as predições de votação estadual dos modelos foram consolidadas em predições de votação nacional, e comparadas com os valores reais de votação nos candidatos da situação entre 2002 e 2018. Nessa fase a precisão da predição foi medida pelo MAE.

A. Resultados

Os hiperparâmetros utilizados para o preditor LinearRegression foram os padrões da plataforma Weka 3.8.5. Nessa parametrização o preditor utiliza o método M5 como seletor de atributos durante o treinamento do modelo, e elimina atributos colineares. Portanto, os atributos presentes nos modelos de regressão foram selecionados como os mais relevantes para a predição. Os modelos de regressão linear multivariada foram construídos para o primeiro e segundo turnos conforme as equações (5) e (6), respectivamente. No modelo de primeiro turno o voto depende da taxa de analfabetismo A, do crescimento percentual da inflação I e das intenções de voto coletadas nas pesquisas eleitorais V. O modelo do segundo turno, por sua vez, representa o voto como dependente do *dummy* de terceiro mandato T, além da taxa de analfabetismo e intenções de voto.

$$Y = 20.5969A - 14.4146I + 19.9547V + 26.5363 \quad (5)$$

$$Y = 20.9509A - 4.0109T + 20.8375V + 33.8414 \quad (6)$$

O preditor REPTree foi aplicado utilizando os hiperparâmetros padrões da plataforma Weka 3.8.5. Dessa forma, o preditor aplica a poda como método de redução de erros, fazendo com que nós que não contribuem com o aumento da precisão sejam iterativamente descartados da árvore. Ainda, essa parametrização não limita a profundidade da árvore.

As métricas de avaliação dos modelos de LinearRegression e REPTree foram coletadas durante a etapa de validação cruzada conforme apresentado na tabela (II).

	Turno	LinearRegression	REPTree
MAE	1º	2,05	2,04
	2º	2,39	2,16
R^2	1º	0,56	0,58
	2º	0,49	0,48

TABLE II
MÉTRICAS DA VALIDAÇÃO CRUZADA COM DADOS DE 2002 A 2014.

Por fim, os modelos foram utilizados na predição das eleições presidenciais entre 2002 e 2018. Nenhum dado da eleição de 2018 foi utilizado na etapa de treinamento. As predições foram apresentadas na tabela (III).

V. ANÁLISE DE RESULTADOS

Os modelos de regressão linear multivariada das equações (5) e (6) apresentaram o voto como dependente da taxa de analfabetismo e da intenção de voto popular no primeiro e segundo turnos. No segundo turno o modelo indica uma leve influência negativa do *dummy* do terceiro mandato na votação do candidato da situação. Os modelos indicam que a taxa de analfabetismo nas regiões é o fator que tem a maior influência positiva na quantidade de votos recebidos pelo candidato da situação. A porcentagem de intenção de voto popular, assim como a taxa de analfabetismo, influenciou positivamente o voto de maneira semelhante nos dois turnos. Por outro lado, o crescimento da inflação teve a maior influência negativa no voto, sendo relevante somente no primeiro turno. O termo intercepto no primeiro turno ($\approx 26,54$) e no segundo ($\approx 33,84$) representa a porcentagem de votos que o candidato da situação receberia se todas as variáveis independentes fossem nulas, ou seja, sem nenhuma influência das variáveis independentes.

Votos do candidato da situação			Predição, em % (Erro, em %)	
Ano	Turno	Real (%)	Linear Regression	REPTree
2002	1º	23,19	20,24 (-2,95)	22,08 (-1,11)
	2º	38,73	38,22 (-0,51)	42,43 (3,70)
2006	1º	48,61	48,66 (0,05)	45,05 (-3,56)
	2º	60,83	61,86 (1,03)	58,27 (-2,56)
2010	1º	46,90	45,75 (-1,15)	46,62 (-0,28)
	2º	56,05	54,10 (-1,95)	58,24 (2,19)
2014	1º	41,59	44,47 (2,88)	43,94 (2,35)
	2º	51,69	52,72 (1,03)	51,36 (-0,33)
2018 [1]	1º	29,28	29,48 (0,20)	41,85 (12,57)
	2º	44,87	41,98 (-2,89)	42,31 (-2,56)
MAE	1º	-	1,45	3,97
	2º	-	1,48	2,27

[1]Predição fora do treinamento.

TABLE III
PREDIÇÕES DAS ELEIÇÕES PRESIDENCIAIS DE 2002 A 2018.

A árvore de regressão gerada no treinamento do primeiro turno apresenta as despesas de campanha do candidato como raiz, e indica o voto como dependente da taxa de analfabetismo, do estado e do crescimento do PIB. No segundo turno, a árvore representa o voto como dependente da taxa de analfabetismo, da intenção de voto popular e dos estados. No primeiro turno, campanhas com despesa máxima menor que 88 milhões de reais (24% da despesa máxima do treinamento) não dependem de nenhuma outra variável. No segundo turno, por outro lado, o voto nos estados com taxa de analfabetismo menor que 11.05% dependem apenas da variável de intenção de voto popular.

O *lead* dos experimentos precisa ser avaliado conforme os turnos do pleito. De forma geral, a disponibilidade da pesquisa de intenção de voto foi a variável limitante para o *lead*. No primeiro turno, o *lead* teve seu mínimo de 24 dias em 2018 e o máximo de 228 dias em 2014 com média de ≈ 110 dias. O

segundo turno historicamente acontece pouco tempo depois do primeiro. Em alguns casos, como na eleição presidencial de 2018, os dois turnos aconteceram no mesmo mês. No segundo turno o *lead* teve seu mínimo de 16 dias em 2002 e máximo de 23 dias em 2010.

A variável de intenção de voto popular apareceu nos dois modelos construídos para o segundo turno. Isso pode ser explicado pelo baixo *lead* comumente observado nesse caso, fazendo com que esse dado já estivesse próximo da votação real.

Durante a validação cruzada ambos os modelos apresentaram R^2 próximos a 0,5, indicando que os modelos são capazes de representar cerca de 50% das variações dos valores reais nos testes. No primeiro turno essa métrica foi mais alta, chegando a 0,58 na árvore de regressão.

O MAE da árvore de regressão foi menor nos dois turnos durante a validação cruzada. Mesmo assim, os resultados dessa métrica indicam um erro médio de $\approx 9\%$ na predição das votações estaduais. Esse erro pode ser considerado alto já que eleições presidenciais como a de 2010, 2014 e 2018 foram decididas com margens menores que essa. No entanto, esse desvio tende a ser mitigado durante a consolidação do voto estadual no nacional.

No teste de predição o modelo de regressão linear multivariada teve menor MAE nos dois turnos, com desvios de predição menores que 1% em 2002, 2006 e 2018. Esse modelo mostrou capacidade de generalização maior que a árvore de regressão evidenciada principalmente na predição da eleição de 2018, que não foi observada no treinamento.

Para realizar uma comparação dos resultados foi selecionado o trabalho realizado por [Turgeon and Rennó 2012]. Nele, os autores construíram um preditor com base em dados apenas de primeiro turno e obtiverem um modelo com 3 variáveis independentes. Na comparação, foi possível observar que o trabalho deles teve R^2 (0,69) maior que os dois modelos construídos pelo presente trabalho para o primeiro turno com o mesmo número de observações. Ainda, o trabalho de [Turgeon and Rennó 2012] foi aplicado nas eleições brasileiras de 1994 a 2010 com MAE $\approx 5,79$ no melhor caso, deixando de fora do treinamento a eleição de 2010, que, por consequência, teve o maior erro de predição. Entretanto, mesmo quando a eleição de 2010 é retirada da análise o modelo de [Turgeon and Rennó 2012] teve MAE $\approx 2,35$. Analisando o trabalho em tela, o modelo de regressão linear treinado para o primeiro turno tem MAE $\approx 1,75$ dentro da amostra de treinamento e $\approx 1,45$ considerando a predição de 2018. Portanto, os modelos treinados no trabalho em tela têm erros de predição menores, dentro e fora da amostra de treinamento, que o melhor modelo apresentado por [Turgeon and Rennó 2012].

O *lead* de 2 meses no primeiro turno apresentado por [Turgeon and Rennó 2012] foi menor que a média de ≈ 110 dias (3,6 meses) utilizada pelo modelo de regressão linear multivariada construído no trabalho em tela.

VI. Conclusão

Esse trabalho abordou a predição de resultados eleitorais utilizando regressão linear multivariada, conforme já amplamente aplicado em democracias consolidadas. No Brasil, essa técnica também foi préviamente utilizada com dados no nível estadual para ampliação do número de observações de treinamento. Ainda, esse trabalho propôs a árvore de regressão como técnica alternativa, apresentando resultados inferiores à regressão linear multivariada.

A seleção de dados seguiu a modelagem proposta por Lewis-Beck [Lewis-Beck 2005] e revelou o impacto relevante da taxa de analfabetismo, que pode ser entendida como uma variável social derivada do cenário econômico, no voto dos dois turnos. A influência da variável de intenção de voto popular nos dois turnos reforça o que foi apresentado em predições anteriores, inclusive em outros países [Jennings et al. 2020], [Lozano and Castillo 2008], [Nadeau et al. 2010]. No entanto, dados relacionados a essa variável foram limitados ao nível nacional por conta da falta de disponibilidade dessa informação no nível estadual, conforme já observado préviamente em eleições brasileiras [Turgeon and Rennó 2012].

O *lead* dos dados utilizados no treinamento foi limitado pela disponibilidade das pesquisas de intenção de voto, principalmente em 2006 e 2018. No entanto, esse *lead* pode ser ainda mais impactado pela disponibilidade de dados econômicos do nível estadual, que podem ser disponibilizados até dois anos após a eleição [Turgeon and Rennó 2012]. Os dados de inflação não impactam o *lead* pois são disponibilizados mensalmente [IBGE 2021d].

Os modelos de regressão linear multivariada apresentaram resultados satisfatórios para a tarefa de predição inclusive fora da amostra de treinamento, apesar da falta de disponibilidade de dados no nível estadual. Todos os modelos apresentaram constantes com sinais coerentes com o conhecimento de domínio. Ainda, essa abordagem possibilitou a construção de modelos de regressão acurados tanto para o primeiro quanto para o segundo turno das eleições. Na comparação com trabalhos relacionados, o modelo de regressão linear multivariada para o primeiro turno apresentou erros de predição menores que o modelo construído pelo trabalho de [Turgeon and Rennó 2012].

Trabalhos futuros podem explorar otimizações de hiperparâmetros que melhorem o desempenho dos modelos de regressão linear multivariada e árvore de regressão. Ainda, outras técnicas de predição de variáveis dependentes contínuas, como o k-NN (*k-nearest neighbors*) ou redes neurais, podem ser aplicadas na mesma base de dados construída nessa pesquisa. Além disso, outros modelos podem ser reconstruídos considerando variáveis como o *marketing* em redes sociais ou o tempo de televisão dos candidatos. Por fim, o recorte de dados foi limitado pela disponibilidade de dados de financiamento de campanha, no entanto, essa variável foi descartada nos dois modelos de regressão linear multivariada. Portanto, novos experimentos podem ser feitos para verificar a influência das eleições de 1994 e 1998 na precisão dos modelos.

REFERENCES

[Brasil 1965] Brasil (1965). Lei nº 4.737/1965. Último acesso em 21 de março de 2021.

[Brasil 1988] Brasil (1988). Constituição da república federativa do brasil. Último acesso em 21 de março de 2021.

[Brasil 2011] Brasil (2011). Lei nº 12.527/2011. Último acesso em 14 de março de 2021.

[Datafolha 2021] Datafolha (2021). Eleições. Último acesso em 05 de junho de 2021.

[Dias 2016] Dias, R. F. (2016). Tancredo Neves e a redemocratização do Brasil. (Portuguese). *Temporalidades*, 7(3):249–274.

[Faceli et al. 2011] Faceli, K., Lorena, A. C., Gama, J., and Carvalho, A. C. P. L. F. D. (2011). *Inteligência Artificial: Uma abordagem de aprendizado de máquina*. LTC.

[Fayyad et al. 1996] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). Knowledge discovery and data mining: Towards a unifying framework. In *KDD-96: Proceedings*, pages 82–88. AAAI Press.

[Frank et al. 2016] Frank, E., Hall, M. A., and Witten, I. H. (2016). *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*. Morgan Kaufmann, 4th edition.

[IBGE 2021a] IBGE (2021a). Pesquisa nacional por amostra de domicílios - pnad. Último acesso em 05 de junho de 2021.

[IBGE 2021b] IBGE (2021b). Produto interno bruto dos municípios. Último acesso em 05 de junho de 2021.

[IBGE 2021c] IBGE (2021c). Síntese de indicadores sociais - sis. Último acesso em 05 de junho de 2021.

[IBGE 2021d] IBGE (2021d). Índice nacional de preços ao consumidor amplo - ipca. Último acesso em 05 de junho de 2021.

[Jennings et al. 2020] Jennings, W., Lewis-Beck, M., and Wlezien, C. (2020). Election forecasting: Too far out? *International Journal of Forecasting*, pages 949–962.

[Kennedy et al. 2017] Kennedy, R., Wojcik, S., and Lazer, D. (2017). Improving election prediction internationally. *Science*, page 515–520.

[Lewis-Beck 2015] Lewis-Beck, M. S. (2015). *Applied Regression: An Introduction (Quantitative Applications in the Social Sciences) Vol 22*. Sage Publications, Inc.

[Lewis-Beck 2005] Lewis-Beck, M. S. (2005). Election forecasting: Principles and practice. *The British Journal of Politics & International Relations*, pages 145–164.

[Lozano and Castillo 2008] Lozano, J. L. S. and Castillo, A. J. (2008). An adaptive model of voting decision: The case of spain. In *XI Applied Economics Meeting*.

[Magalhães et al. 2012] Magalhães, P. C., Aguiar-Conraria, L., and Lewis-Beck, M. S. (2012). Forecasting spanish elections. *International Journal of Forecasting*, pages 769–776.

[Marsland 2009] Marsland, S. (2009). *Machine Learning: an algorithmic perspective*. Boca Raton, Fl: CRCPress, 2nd edition.

[Nadeau et al. 2010] Nadeau, R., Lewis-Beck, M. S., and Éric Bélanger (2010). Electoral forecasting in france: A multi-equation solution. *International Journal of Forecasting*, pages 11–18.

[Russell and Norvig 2004] Russell, S. J. and Norvig, P. (2004). *Inteligência artificial*. Elsevier.

[TSE 2021] TSE (2021). Repositório de dados eleitorais. Último acesso em 13 de março de 2021.

[Turgeon and Rennó 2012] Turgeon, M. and Rennó, L. (2012). Forecasting brazilian presidential elections: Solving the n problem. *International Journal of Forecasting*, pages 804–812.

[Weisberg 2005] Weisberg, S. (2005). *Applied Linear Regression*. WILEY, 3rd edition.

[Witten et al. 2007] Witten, I. H., Frank, E., and Hall, M. A. (2007). *Data mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2nd edition.