Aplicação Combinada de Métodos Baseados em Árvores e Importância da Variável para Análise de Despesas de Campanhas Eleitorais

Alessandra S. Gomes[1]

[1]InternetLab
São Paulo – SP – Brasil

alessandra.gomes@internetlab.org.br

Abstract. *This paper presents the use of tree-based methods and permutation variable importance as an alternative to linear methods for the analysis of electoral campaign spending. Therefore, we analyzed the spending of candidates for mayor in the 2020 elections with ERT algorithm and permutation importance. The results showed that the most important spending were those related to communication and advertising.*

Resumo. *Este artigo apresenta o uso de métodos baseados em árvores e importância da variável por permutação como alternativa aos métodos lineares para a análise de despesas de campanhas eleitorais. Para tanto, analisou-se as despesas de candidatos à prefeitura nas eleições de 2020 com algoritmo ERT e a importância por permutação. Os resultados apontaram que gastos de maior importância foram os relacionados a comunicação e publicidade.*

1. Introdução e Justificativa

A relação entre dinheiro e eleições é um dos temas de pesquisa da Ciência Política. Dentre as frentes investigadas estão descobrir e analisar correlações entre o total gasto em campanhas eleitorais e a quantidade de votos obtidos, assim como investigar se a forma como se gasta o dinheiro influencia no resultado eleitoral.

Várias destas pesquisas fazem uso de métodos lineares para testar suas hipóteses e encontrar seus resultados. Entretanto, a aplicação válida desses métodos exige o atendimento de um conjunto rígido de pressupostos e isto nem sempre é possível, pois dados do mundo real nem sempre são lineares. Dessa forma, neste trabalho será proposto o uso combinado de métodos não paramétricos baseados em árvores e importância da variável como uma alternativa mais flexível para estas análises.

Para aplicação, foi utilizado o algoritmo Extremely Randomized Trees e a função Permutation Variable Importance em um *dataset* contendo dados das eleições de 2020 para a prefeitura dos municípios do Brasil, com o objetivo de encontrar, por meio de uma tarefa de classificação, quais despesas se mostraram mais importantes na diferenciação entre os candidatos eleitos e não eleitos.

Este artigo está organizado da seguinte forma: na Seção 2 são apresentados os principais conceitos sobre a Importância da Variável e Métodos Baseados em Árvores que formam a Fundamentação Teórica deste trabalho. Na Seção 3 é apresentada a Revisão da Literatura relacionada a Análise de Gastos em Despesas de Campanhas Eleitorais e Importância da Variável em Estudos Eleitorais. Na Seção 4 é apresentada a

proposta deste artigo. Na Seção 5 são apresentados e discutidos os resultados obtidos e, por fim, na Seção 6 é apresenta a conclusão e sugestões de trabalhos futuros.

2. Fundamentação Teórica

Nesta seção são apresentadas e desenvolvidas as fundamentações teóricas que embasam a proposta deste trabalho.

2.1. Importância da Variável e Métodos Baseados em Árvores

Importância da variável, do inglês Variable Importance, é uma técnica que calcula o quanto um modelo preditivo depende de cada uma das variáveis que o gerou para realizar predições acuradas [Loh and Zhou 2021]. Logo, variáveis que possuem forte relação com a resposta preditiva serão consideradas preditores importantes e receberão valores altos e/ou positivos e variáveis que se mostrarem preditores fracos receberam valores baixos e/ou negativos. É possível também que variáveis não causem nenhum impacto no modelo e, assim, recebam valor de importância igual a zero.

Uma das formas de se obter a importância das variáveis é por meio de métodos algorítmicos baseados em árvores. Trata-se de uma classe de algoritmos de Aprendizado de Máquina supervisionados e não paramétricos capazes de executar tarefas de classificação e regressão. A proposta foi apresentada em 1984 por Leo Breiman, um estatístico da Universidade da Califórnia, Berkeley [Breiman 1984] e consistia em representar os padrões de um conjunto de dados em formato de árvore, onde os nós representariam as variáveis deste *dataset* e as ramificações seriam construídas a partir de testes condicionais. A implementação desta proposta, chamada de Árvore de Decisão, utiliza a Diminuição Média na Impureza, do inglês Mean Decrease in Impurity (MDI), como teste condicional.

Para se obter a Árvore de Decisão que melhor representa os dados é necessário calcular qual variável é a mais indicada como nó raiz e qual valor de *threshold* irá maximizar o critério de divisão da ramificação. Dessa forma, nota-se que a construção deste modelo já possui um componente que indica o grau de importância das variáveis.

De uma forma geral, Árvores de Decisão geram modelos preditivos bons e de fácil interpretação. Entretanto, possuem alta variância e fácil tendência ao *overfitting*, o que pode ser prejudicial para a análise de importância de variável. A solução para estes problemas foi proposta por Breiman em 1996 com a utilização de combinação de várias árvores de decisão para alcançar resultados melhores [Breiman 1996]. Esta técnica é conhecida como métodos ensemble e os mais conhecidos são o *boosting* e o *bagging*. No primeiro caso, as árvores são treinadas sequencialmente e gradualmente, de forma que o treinamento da árvore posterior busca ajustar os erros de treinamento ocorridos na árvore anterior. No segundo caso, as árvores são treinadas isoladamente e a decisão final do modelo é realizada por meio de votação ou média.

A melhoria nos resultados proporcionada pelos métodos ensemble tem por consequência melhor qualidade nos resultados de importância da variável. Alguns algoritmos que implementam o método *boosting* são AdaBoost e o Gradient Boosting. Dos que implementam o método *bagging* pode-se citar o Random Forest, também proposto por Breiman em 2001 [Breiman 2001] e o Extremely Randomized Trees, que será o utilizado neste artigo.

2.2. Injeção de Aleatoriedade e Extremely Randomized Trees

A injeção de aleatoriedade em algoritmos baseados em árvores permite a construção de modelos melhores, pois possibilita que as árvores possuam mais diversidade entre si. Dessa forma, a geração do modelo sofrerá menos influência de correlações indesejadas. Como consequência, este modelo possuirá melhor capacidade de generalização, pois será capaz de realizar predições para diferentes possibilidades e, assim, alcançará melhor desempenho [Liu 2005]. E, como apontado previamente, a qualidade da estimação da importância das variáveis está fortemente relacionada com a qualidade do modelo gerado.

Assim, optou-se por utilizar neste trabalho o algoritmo Extremely Randomized Trees (ERT) dado os vários fatores de aleatoriedade existentes em sua implementação. Publicado em 2006, este algoritmo adota a aleatoriedade na seleção de amostras, com reposição, para o treinamento das árvores do modelo, na seleção de variáveis como nó raiz de cada árvore e na seleção do valor inicial de *threshold* para o *split* das observações para cada árvore. A utilização do cálculo MDI se inicia a partir deste ponto, para dar continuidade a ramificação de cada árvore [Geurts and Ernst and Wehenkel 2006].

Apesar de todas as vantagens proporcionadas pela aleatoriedade, ela não é capaz de solucionar o dilema viés-variância, ou seja, inevitavelmente, a redução da variância, que implica na melhoria da acurácia, terá como consequência o aumento de viés no modelo, que implica em perda de precisão [Doroudi 2020]. Por este motivo, a análise de importância de variável será combinada com a técnica Importância da Variável por Permutação, do inglês Permutation Variable Importance.

2.3. Redução de viés e Importância da Variável por Permutação

O custo da inserção de aleatoriedade é o aumento do viés, que pode levar o modelo a dar um alto valor de importância para variáveis que não são tão importantes para a tarefa da predição. Uma alternativa para lidar com este problema é a utilização da Importância da Variável por Permutação. Este recurso avalia a importância de uma variável alterando a informação ali contida. Caso essa alteração cause impacto na predição, então a variável possuía informação valiosa para o modelo. Assim, quanto maior o impacto, mais importante esta variável de fato é para o modelo.

A alteração da informação é realizada por meio de embaralhamento aleatório e o cálculo da importância da variável é feito utilizando a distância entre o resultado da predição após o embaralhamento e a predição original. Logo, quanto maior o decréscimo na qualidade de predição, maior a importância daquela variável para o modelo. Esta técnica é considerada agnóstica, pois pode ser aplicada a diferentes tipos de modelo de aprendizado de máquina, não apenas os baseados em árvores.

3. Revisão da Literatura

Nesta seção será apresentada uma breve revisão da literatura vinculada a trabalhos eleitorais correlatos que analisaram diferentes períodos e corridas eleitorais.

3.1. Análise de Gastos em Despesas de Campanhas Eleitorais

A literatura brasileira de Ciência Política possui diversos estudos que investigam se a forma como o dinheiro é gasto durante as campanhas eleitorais pode influenciar no resultado das eleições. Muitos destes trabalhos utilizaram diferentes métodos lineares

para testar as hipóteses e analisar os resultados obtidos. Em [Paranhos 2017] os autores utilizam regressão linear para analisar a relação entre os tipos de gastos realizados por candidatos à prefeitura durante a campanha eleitoral dos anos 2008, 2012 e 2016 e o total de votos obtidos. Em [Paranhos 2018] os autores utilizam técnicas de análise multivariada para analisar a relação entre um conjunto pré-definido de tipos de despesas eleitorais e a conquista de votos nas eleições de 2008, 2012 e 2016 para a prefeitura.

Em alguns trabalhos, os pesquisadores utilizam agrupamento de tipos de despesas para fazer análises mais agregadas. Em [Sampaio 2021] os autores agrupam as despesas em três categorias e utilizam regressão linear e análise descritiva para analisar a relação delas com a quantidade de votos obtidos nas eleições municipais de 2016. Em [Speck and Mancuso 2017] os autores adotam a categorização despesas tradicionais e modernas e utilizam regressão logística para analisar qual categoria resulta em maior sucesso nas urnas para os candidatos a cargos proporcionais e majoritários nas eleições de 2014.

Por fim, em outros trabalhos, os pesquisadores fazem análises mais específicas e investigam a importância de um pequeno conjunto de tipos de despesas para o resultado eleitoral. Em [Speck and Cervi 2016] a regressão linear múltipla é utilizada para analisar a relação entre um conjunto específico de variáveis que, além de gastos, também representam tempo de propaganda eleitoral e capital político, e o resultado eleitoral para candidatos à prefeitura em 2012. Em [Castro and Viana 2018] a regressão multivariada é utilizada para analisar o impacto que gastos em despesas de TICs tiveram sobre o resultado das candidaturas à prefeitura em 2016. Em [Heiler and Viana and Santos 2016] os autores utilizam regressão logística multivariada para analisar o impacto de gastos em estrutura e comunicação no sucesso eleitoral para candidatos a deputado federal em 2010 e se estratégias de gastos concentrados em poucos tipos de gastos teriam menor probabilidade de obter vitória nas urnas.

3.2. Importância da Variável em Estudos Eleitorais

Pesquisas que utilizam importância de variáveis ou outras técnicas diferentes de regressão são encontradas em trabalhos sobre estudos eleitorais, porém são mais frequentes em publicações internacionais e não relacionadas especificamente a análise de gastos. Em [Deveci and Keser 2020] os autores utilizaram importância da variável para encontrar as variáveis mais importantes para a decisão de voto dos eleitores em eleições municipais na Turquia. Os pesquisadores identificaram que as variáveis mais importantes são o resultado de eleições anteriores, o valor médio de moradores de uma residência e a proporção da população jovem dos municípios. Em [Dey and Alvarez 2021] os autores utilizaram importância da variável para analisar o resultado das eleições de 2020 para presidência dos EUA. Neste trabalho, as variáveis mais importantes encontradas foram a identificação e a polarização partidária.

4. Implementação da Proposta

Neste trabalho, serão utilizados os dados das eleições de 2020 para analisar quais despesas foram mais importantes na diferenciação entre candidatos eleitos e não eleitos. Nestas eleições ocorreram as corridas para prefeitos e vereadores. A primeira é decidida de acordo com o sistema majoritário, ou seja, o candidato eleito é o que obtém a maioria dos votos. A segunda é decidida de acordo com o sistema proporcional, onde se aplica o cálculo do quociente eleitoral para decidir se um candidato foi ou não eleito [TRE-SC 2022].

Optou-se então por analisar apenas os dados relacionados à eleição majoritária, pois nela a relação gastos e resultado eleitoral é mais direta do que no sistema proporcional, onde outras variáveis são consideradas no processo. Assim, aqui serão analisadas apenas as despesas dos candidatos à prefeitura. Por fim, optou-se também analisar apenas os dados dos candidatos cujos municípios elegeram seus prefeitos no primeiro turno. Esse escopo foi definido, pois os candidatos que foram para o segundo turno tiveram um período de tempo maior para efetuar gastos.

As implementações realizadas nas etapas seguintes foram feitas utilizando a linguagem Python, a biblioteca de análise de dados Pandas e o *framework open-source* H2O para o uso de algoritmos de Aprendizado de Máquina. Os gráficos apresentados nesta seção foram gerados com Tableau Public. A implementação completa está disponível em https://github.com/alegomesbr/wcge2022.

4.1. Coleta e Pré-processamento dos Dados

O Tribunal Superior Eleitoral (TSE) disponibiliza os dados relacionados às eleições para consulta e download em seu portal de Dados Abertos [TSE 2022]. Nele, é possível obter dados a partir das eleições de 1994. No formato para download, os dados são organizados por ano eleitoral e em três diferentes arquivos CSV. Um contém os dados sobre todos os candidatos que registraram candidatura, outro contém dados sobre todos os votos obtidos pelos candidatos e o último contém os dados sobre as despesas de campanha eleitoral declaradas pelos candidatos.

Para este trabalho, foi criado um *dataset* a partir do cruzamento de dois destes arquivos: o que possui os registros dos candidatos e o que possui os dados das despesas declaradas. Os atributos selecionados nesta primeira etapa do pré-processamento dos dados são apresentados na Tabela 1.

Table 1. Campos (Atributos) selecionadas a partir do cruzamento dos *datasets*

Campos	**Descrição**	***Dataset* de Origem**
`sq_candidato`	Identificador único dos candidatos	Dados sobre os Candidatos Dados sobre as Despesas
`ds_situacao_candidatura`	Identifica se a(o) candidata(o) está apta ou inapta para concorrer às eleições	Dados sobre os Candidatos
`ds_turno`	Identificação do turno	Dados sobre os Candidatos
`ds_municipio`	Município para onde a(o) candidata(o) irá concorrer	Dados sobre os Candidatos
`ds_cargo`	Cargo para qual a(o) candidata(o) irá concorrer	Dados sobre os Candidatos
`ds_sit_tot_turno`	Identifica se a(o) candidata(o) foi eleito, não eleito ou se foi para o segundo turno	Dados sobre os Candidatos
`ds_origem_despesa`	Nome da categoria da despesa cujo valor foi declarado	Dados sobre as Despesas
`ds_valor_despesa`	Valor declarado para uma respectiva categoria de despesa	Dados sobre as Despesas

Dois filtros foram aplicados no *dataset* resultante deste cruzamento. O primeiro selecionou todos os candidatos que se encontravam aptos e que concorriam ao cargo de prefeito. O segundo selecionou todos os municípios cujos candidatos à prefeitura foram eleitos no primeiro turno. Após a execução dos filtros, os campos

`ds_situacao_candidatura`, `ds_turno`, `ds_municipio` e `ds_cargo` foram excluídos, pois seria redundante mantê-los. O *dataset* final resultou em 17.053 candidatos que concorreram às eleições em 5.414 municípios do país.

A soma total aproximada das despesas declaradas foi de R$1.5B. Estes gastos foram distribuídos por 41 categorias, definidas pelo TSE para identificar a origem dos gastos declarados (`ds_origem_despesa`). De acordo com a classificação apresentada em [Heiler and Viana and Santos 2016], as despesas eleitorais nas eleições de 2020 foram:

Gasto com pessoal: Água; Alimentação; Atividades de militância e mobilização de rua; Encargos sociais; Despesas com pessoal; Diversas a especificar; Doações financeiras a outros candidatos e/ou partidos

Comunicação e publicidade: Comícios; Eventos de promoção da candidatura; Produção de jingles, vinhetas e slogans; Publicidade por carros de som; Publicidade por jornais e revistas; Publicidade por materiais impressos; Criação e inclusão de páginas na internet; Despesa com Impulsionamento de Conteúdos; Correspondências e despesas postais; Publicidade por adesivos; Produção de programas de rádio televisão ou vídeo; Telefone

Estrutura: Cessão ou locação de veículos; Combustíveis e lubrificantes; Despesas com transporte ou deslocamento; Despesas com Hospedagem; Passagem Aérea; Encargos financeiros, taxas bancárias e ou op. cartão de crédito; Energia elétrica; Impostos, contribuições e taxas; Locação, cessão de bens imóveis; Locação, cessão de bens móveis exceto veículos; Materiais de expediente; Multas eleitorais; Pesquisas ou testes eleitorais; Reembolsos de gastos realizados por eleitores; Despesa com geradores de energia; Pré-instalação física de comitê de campanha; Aquisição, Doação de bens móveis ou imóveis; Taxa de Administração de Financiamento Coletivo; Serviços advocatícios; Serviços contábeis; Serviços prestados por terceiros; Serviços próprios prestados por terceiros

A última etapa de preparação do *dataset* foi a conversão de todos os valores de gastos para seus respectivos valores percentuais, considerando como total a soma de todos os gastos de cada candidato. Esta conversão foi feita, pois neste trabalho o objetivo é contribuir com a análise de como os candidatos eleitos e não eleitos distribuíram o total gasto durante a campanha eleitoral pelas despesas existentes. Com esta conversão, é possível analisar quanto do valor percentual foi distribuído para cada tipo de despesa.

4.2. Implementação da Importância da Variável

Com o *dataset* pronto, a próxima etapa foi a implementação do algoritmo ERT com o recurso da Importância da Variável por Permutação. Para tanto, foi utilizada a linguagem Python e o *framework* H2O para a escrita dos códigos relacionados a Aprendizado de Máquina. Os gráficos apresentados nesta seção foram gerados no Tableau Public. Assim, o pseudo-código abaixo apresenta a lógica do código criado:

```
1. Carregamento dos Dados
2. Definição das variáveis preditoras (x) e resposta (y)
3. Definição dos conjuntos de treinamento, teste e validação
4. Construção do modelo
5. Treinamento e validação do modelo
6. Cálculo de métricas para avaliar o modelo
7. Aplicação de Importância da Variável por Permutação
```

Nos Passos 1 e 2 o *dataset* criado é carregado e os campos `ds_origem_despesa` e `ds_valor_despesa` são atribuídos como variáveis preditoras e `ds_sit_tot_turno` como variável resposta, cujas respostas para a tarefa de classificação binária serão `ELEITO` e `NÃO ELEITO`. O campo `sq_candidato` é utilizado apenas como identificador único dos candidatos. Em seguida, os conjuntos de treinamento, teste e validação são definidos, aleatoriamente, na proporção 70%, 15% e 15%, respectivamente (Passo 3).

No Passo 4 o modelo é construído. Para tanto, foi utilizada a classe `H2ORandomForestEstimator` do *framework* H2O. Além de configurações de Tuning, destacam-se a configuração dos parâmetros `histogram_type = Random` e `balance_classes = True`. O primeiro identifica que o algoritmo a ser implementado será o ERT. O segundo aponta para o algoritmo que técnicas de balanceamento de dados devem ser aplicadas, pois o *dataset* em questão é desbalanceado. A decisão de balancear os dados na etapa da implementação e não da manipulação dos dados se deve ao fato do contexto ser naturalmente desbalanceado, ou seja, em todo processo eleitoral a quantidade de candidatos não eleitos será muito maior que a de eleitos. Nas eleições de 2020, a proporção para candidaturas à prefeitura foi de 31.75% eleitos e 68.25% não eleitos. Uma manipulação direta no *dataset* com exclusão ou inclusão de novos registros iria descaracterizar o evento já ocorrido.

Com o modelo configurado, os dois passos seguintes são a realização do treinamento e validação e o cálculo de métricas para avaliar o desempenho do modelo (Passos 5 e 6). Por fim, o último passo é a execução da Importância da Variável por Permutação para se obter o ranking das variáveis mais importantes para o modelo (Passo 7). A implementação deste recurso foi feita com o uso da função `varimp()`, também do *framework* H2O. Seguindo a recomendação indicada pela documentação do próprio *framework*, os dados utilizados nesta etapa foram os definidos para o conjunto de teste [Molnar 2022].

5. Resultados e Discussões

Nesta seção serão apresentados os resultados obtidos pelo modelo gerado e possíveis interpretações sobre as despesas consideradas mais importantes e a forma não concentrada de gastos duranta a campanha eleitoral.

Devido a natureza randômica da distribuição do *dataset* em conjuntos de treinamento, validação e teste, e do algoritmo ERT, a implementação foi executada 50 vezes. Assim, os resultados aqui apresentados representam a média destas execuções.

5.1. Avaliação do Modelo

Considerando que o modelo foi treinado com um *dataset* desbalanceado, optou-se por utilizar as métricas mais indicadas para situações de desbalanceamento de dados [Haibo and Yunqian 2013]. Logo, foram utilizados o Macro-Averaged Accuracy (MAA), Geometric Mean (G-Mean) e F-Measure para avaliar a performance do modelo. O MAA, para o caso de classificações binárias, é a média aritmética entre a acurácia de verdadeiros positivos (*sensitivity*) e de verdadeiros negativos (*specificity*). Logo, é uma métrica que calcula a média aritmética entre as acurácias de cada classe.

O G-Mean calcula o desempenho relativo do modelo para a classificação das classes positivas e negativas. Um valor baixo de G-Mean indica que o modelo não está tendo um bom desempenho na classificação de uma das classes. Por fim, o F-Measure é a média harmônica entre precisão e *recall*. Logo um baixo valor de F-Measure indica que há uma presença significativa de falsos positivos ou falsos negativos. As métricas e os respectivos valores obtidos são apresentados na Tabela 2.

Table 2. Métricas e valores obtidos para o modelo gerado com ERT

F-Measure	MAA	G-Mean
0.7136	0.6391	0.6115

5.2. Análise da Importância das Variáveis

O *ranking* as variáveis mais importantes para o modelo gerado de acordo com o percentual de contribuição na predição entre eleitos e não eleitos é apresentada na Figura 1:

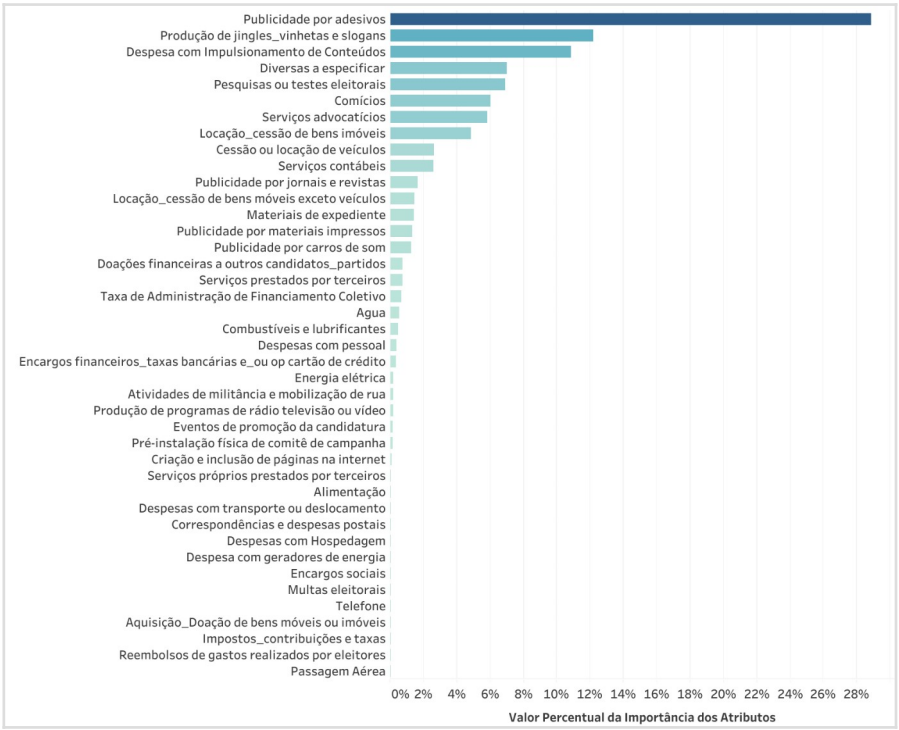

Figura 1. *Ranking* das variáveis mais importantes para a classificação de candidatos à prefeitura em eleitos ou não eleitos nas eleições de 2020

Pelo gráfico da Figura 1, pode-se notar que a despesa “Publicidade por adesivos”

se destaca das demais em termos de importância. Ela é responsável por contribuir com aproximadamente 28.89% da decisão da tarefa de predição pelo modelo gerado. As duas despesas seguintes são “Produção de jingles, vinhetas e slogans”, com 12.16% e “Despesa com Impulsionamento de Conteúdo” com 10.85%. Pode-se notar aqui também que estas são as únicas despesas que possuem importância acima de 10%.

Na faixa de contribuições entre 5% e 10% estão as despesas “Diversas a especificar”, com 6.98%, “Pesquisas ou testes eleitorais” com 6.91%, “Comícios” com 6.04% e “Serviços advocatícios” com 5.82%. Entre 2% e 5% temos 8 despesas: “Locação ou cessão de bens imóveis” com 4.84%, “Cessão ou locação de veículos” com 2.63%, Serviços contábeis com 2.58%, “Publicidade por jornais e revistas” com 1.65%, “Locação ou cessão de bens móveis exceto veículos” com 1.46%, “Materiais de expediente” com 1.43%, “Publicidade por materiais impressos” com 1.30% e “Publicidade por carro de som” com 1.29%.

Por fim, as demais 26 despesas estão abaixo de 1% de importância. Destas, as despesas “Impostos, contribuições e taxas”, “Reembolso de gastos realizados por eleitores” e “Passagem Aérea” retornaram 0%, ou seja, não causaram nenhum impacto no modelo.

Pode-se notar que as três despesas consideradas mais importantes pertencem a categoria Comunicação e Publicidade e somam juntas aproximadamente 52% das contribuições para a decisão do modelo. Logo, um pouco mais de 50% das contribuições para a predição do modelo em classificar um candidato como eleito ou não eleito são baseadas nos investimentos feitos nas diferentes formas dos candidatos se comunicarem com os eleitores ou divulgarem publicitariamente suas propostas. E os meios destas três categorias são adesivos, jingles, vinhetas, slogans e/ou redes sociais.

Para enriquecer a análise das três despesas mais importantes, a Figura 2 resume uma breve análise exploratória sobre o percentual de candidatos que investiram e não investiram nessas 3 despesas e a Figura 3 apresenta as curvas de Dependência Parcial de cada despesa em relação a variável resposta, ou seja, a situação final do candidato (`ELEITO` ou `NÃO ELEITO`).

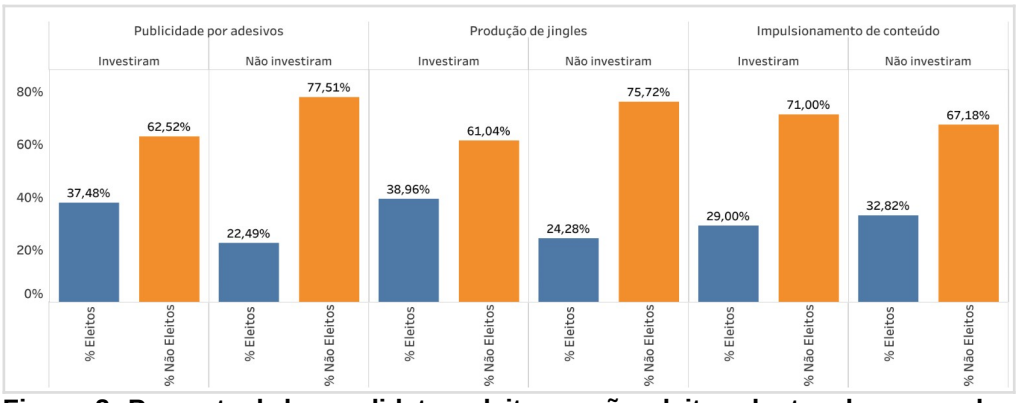

Figura 2. Percentual de candidatos eleitos e não eleitos dentro do grupo dos que investiram e não investiram nas três categorias de despesas consideradas mais importantes pelo modelo

Pela Figura 2 podemos observar a presença percentual de candidatos eleitos e não eleitos para cada uma das 3 despesas de acordo com a declaração de gastos. Ao comparar os candidatos que declararam e não declararam gastos com “Publicidade por

adesivos", podemos observar que, percentualmente, a presença de candidatos eleitos foi maior no grupo dos que declararam esta despesa. O mesmo ocorreu para "Produção de jingles, vinhetas e slogans". Em "Despesas com Impulsionamento de Conteúdo" ocorreu o contrário, com maior presença percentual de eleitos no grupo dos que não declararam esta despesa.

Os Gráficos de Dependência Parcial, do inglês Partial Dependence Plot (PDP), são um recurso que mostram visualmente a relação de dependência parcial entre o modelo e uma determinada variável. No gráfico da Figura 3, o eixo x representa o percentual de gastos destinados a cada uma das três despesas e o eixo y a probabilidade média do candidato ser classificado como ELEITO. Em uma rápida análise é possível perceber que a relação entre as 3 despesas e a variável resposta é não linear e com algum tipo de variação entre 0% e aproximadamente 20%. A partir de 20% as três curvas possuem comportamento constante.

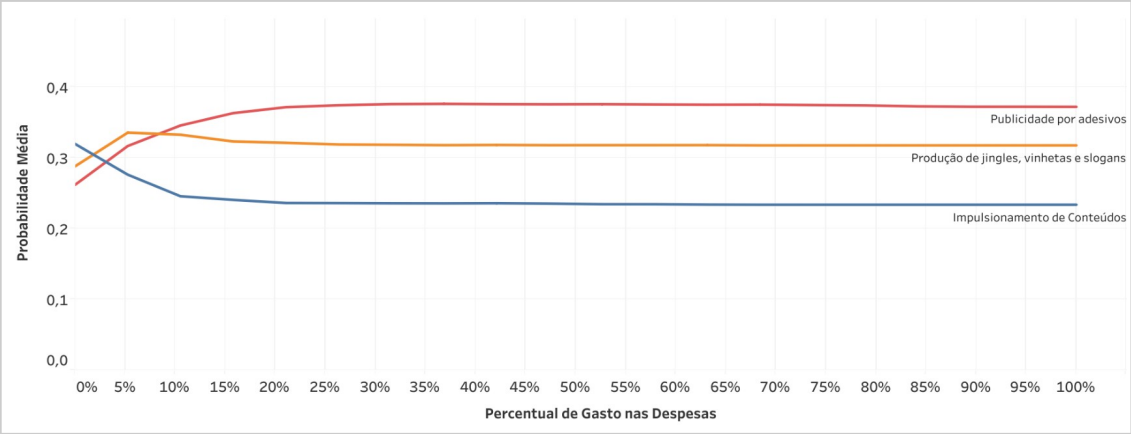

Figura 3. Gráfico de Dependência Parcial para as três despesas mais importantes para o modelo

A análise exploratória apresentada na Figura 2 referente a "Publicidade por adesivos" e "Produção de jingles, vinhetas e slogans" dá um indicativo de que investimento nestas despesas possui forte contribuição para classificar um candidato como ELEITO. Analisando as curvas na Figura 3 para "Publicidade por adesivos" a curva possui comportamento positivo entre 0% e 20%, ou seja, a probabilidade de ser classificado ELEITO pelo modelo cresce na medida em que o percentual de investimento nesta despesa cresce até 20%. Para "Produção de jingles, vinhetas e slogans" o comportamento entre 0% e 20% é um pouco mais complexo, com a probabilidade média de ser considerado eleito atingindo seu pico em aproximadamente 5% e, a partir de então, segue em decrescimento até aproximadamente 20% ou 25%.

Ao contrário das outras duas despesas, a análise exploratória para "Despesas com Impulsionamento de Conteúdo" (Figura 2) forne indícios de que investimento nesta despesa possui contribuição para classificar um candidato como NÃO ELEITO. Pelo gráfico PDP (Figura 3), pode-se observar que a relação de dependência parcial desta despesa com a variável resposta possui comportamento negativo, ou seja, a probabilidade média de ser considerado ELEITO decresce na medida em que se aumenta o investimento nesta despesa até aproximadamente 20%.

As variações existentes entre 0% e 20% nestas despesas não indicam uma relação de causalidade direta, onde o candidato que investir menos de 20% em eleições

futuras terá maior probabilidade de ser eleito. Porém, elas refletem a estratégia de distribuir os gastos em diferentes tipos de despesas que pesquisadores de Ciências Políticas já apontaram como tendo maior probabilidade de obter sucesso nas urnas [Heiler and Viana and Santos 2016]. Assim, nas eleições de 2020, investimentos abaixo de 20% do gasto total em diferentes meios de comunicação e publicidade se mostraram mais vencedores do que um alto investimento concentrado em um único meio de comunicação e publicidade.

6. Conclusão e Trabalhos Futuros

Este trabalho apresentou a proposta de uso de métodos não paramétricos baseados em árvores como alternativa aos métodos lineares para a identificação das despesas eleitorais mais importantes durante a corrida eleitoral para prefeitura de 2020. Para tanto, foi gerado um modelo de aprendizado de máquina com o algoritmo ERT e calculado a importância das variáveis com a técnica de importância de variável por permutação. Com o auxílio de uma breve análise exploratória e análise de dependência parcial, identificou-se que as três despesas mais importantes estavam diretamente relacionadas à comunicação e publicidade e que a estratégia de baixos investimentos em diferentes tipos de despesas mostrou ter maior probabilidade média de sucesso eleitoral.

Os trabalhos futuros podem explorar novos recortes, como gênero, raça e regionalidade, e assim, observar o quanto os resultados gerais aqui apresentados sofrem modificações. Além disso, pode-se também investigar hipóteses existentes na Ciência Política como a relação entre gastos concentrados ou dispersos e o resultado eleitoral ou ainda quais despesas possuem mais importância para candidatos veteranos ou iniciantes.

Um outro desafio é investigar quais tecnologias de Aprendizado de Máquina seriam mais adequadas para se utilizar nos casos de eleições não majoritárias, ou seja, as proporcionais, visto que o resultado desta vai além da contabilidade dos votos.

Referências

Breiman, L. et al (1984). “Classification And Regression Trees”. 1st ed. Routledge. https://doi.org/10.1201/9781315139470

Breiman, L. (1996). “Bagging Predictors”. Machine Learning, 24, 123-140.

Breiman, L. (2001). “Random Forests”. Machine Learning. 45. 5-32. 10.1023/A:1010950718922.

Castro, P. A. B. and Viana, F. M. (2018). "Despesas de campanha e retorno eleitoral dos candidatos a prefeito: estratégias tradicionais e uso de TICs nas eleições municipais de 2016". In: ENCONTRO ANUAL DA ANPOCS, Caxambu, MG.

Deveci, K. I. and Keser, I. (2020). "Exploring Decision Rules for Election Results by Classification Trees". KnE Social Sciences.

Dey, S. and Alvarez, R. M. (2021). "Fuzzy Forests For Feature Selection in High-Dimensional Survey Data: An Application to the 2020 U.S. Presidential Election". In: The 3rd International Conference on Applied Machine Learning and Data Analytics

Doroudi, S. (2020). "The Bias-Variance Tradeoff: How Data Science Can Inform Educational Debates". AERA Open. https://doi.org/10.1177/2332858420977208

Geurts, P. and Ernst, D. and Wehenkel, L. (2006). "Extremely randomized trees". Mach

Learn 63, 3–42. https://doi.org/10.1007/s10994-006-6226-1

Haibo, H. and Yunqian, M. (2013). "Assessment Metrics for Imbalanced Learning," in Imbalanced Learning: Foundations, Algorithms, and Applications, IEEE, pp.187-206

Heiler, J. G. and Viana, J. P. S. L. and Santos, R. D. (2016) "O custo da política subnacional: a forma como o dinheiro é gasto importa? Relação entre receita, despesas e sucesso eleitoral". Opinião Pública [online]. v. 22, n. 1. https://doi.org/10.1590/1807-0191201622156. Março.

Liu, F. T. (2005). "The Utility of Randomness in Decision Tree Ensembles". https://feitonyliu.files.wordpress.com/2009/08/master.pdf, Abril.

Loh, W., and Zhou, P. (2021). "Variable Importance Scores". In: Journal of Data Science, 19(4), 569-592. doi:10.6339/21-JDS1023

Molnar, C. (2022). "Interpretable Machine Learning: A Guide for Making Black Box Models Explainable" 2nd ed. christophm.github.io/interpretable-ml-book/

Paranhos, R. et al. (2017) "Meu dinheiro, minhas regras: gastos de campanhas em eleições para prefeitos no Brasil (2008-2016)". In: CONGRESSO LATINO-AMERICANO DE CIÊNCIA POLITICA, Montevidéu, Uruguai.

Paranhos, R. et al. (2018) "Gastos de Campanha nas Eleições Municipais (2008-2016): Uma análise dos tipos de gastos de campanha". In: XI ENCONTRO DA ABPCP, Curitiba, PR.

Sampaio, D. (2021). "CAMPANHAS TRADICIONAIS OU MODERNAS? Estratégias de gastos nas eleições municipais de 2016". In: Revista Brasileira de Ciências Sociais [online]. 2021, v. 36, n. 105. https://doi.org/10.1590/3610511/2020. Abril.

Speck, B. W. and Cervi, E. U. (2016). "O peso do dinheiro e do tempo de rádio e TV na disputa do voto para prefeito". In . Rio de Janeiro: FGV.

Speck, B. W. and Mancuso, W. P. (2017), "'Street fighters' e 'media stars': estratégias de campanha e sua eficácia nas eleições brasileiras de 2014". In: Cadernos Adenauer, v. 3, nº 7, pp. 121-138.

TRE-SC. (2022). "Eleições majoritárias e proporcionais". https://apps.tre-sc.jus.br/site/eleicoes/eleicoes-majoritarias-e-proporcionais. Abril.

TSE. (2022) Portal de Dados Abertos do TSE. https://dadosabertos.tse.jus.br/dataset/. Abril.