Gabinete Compartilhado.

Laranjômetro - Previsão de candidaturas laranjas em 2020

Nota Técnica nº 003/2020

Novembro de 2020

Introdução

Desde 2009, a legislação eleitoral exige que, do total das candidaturas apresentadas por uma coligação em uma eleição proporcional, um mínimo de 30% deve ser reservada a cada sexo. Na grande maioria dos casos, isso se traduz numa cota mínima de 30% de candidaturas femininas, dado que a ampla maioria das candidaturas apresentadas pelas coligações tende a ser de homens.

Para poder apresentar o máximo possível de candidaturas masculinas, as coligações sem um número suficiente de candidatas mulheres fazem uso das **candidaturas laranjas:** são mulheres que apenas se inscrevem (ou são inscritas) como candidatas (cumprindo formalmente a cota) mas não fazem campanha e acabam, muitas vezes obtendo zero votos.

O projeto "Laranjômetro" consistiu em identificar, antes da votação das eleições municipais de 2020:

- candidaturas aos cargos de vereador com risco de serem laranjas;
- coligações que violam ou que apresentam riscos de violar a cota feminina, seja por não atingirem o mínimo de 30% de candidaturas femininas ou por atingirem o mínimo através do uso de candidaturas laranjas.

Para tanto, utilizamos dados das eleições proporcionais brasileiras desde 2004, modelos de aprendizagem de máquina (*machine learning*) e outros métodos de análise de dados. Em linhas gerais, esse projeto se dividiu nas seguintes etapas:

1. Identificação, nos dados disponíveis, de um *proxy*[1] para candidaturas laranjas;
2. Construção de características (*features*) das candidaturas com potencial de distinguir candidaturas laranjas das demais (denominadas regulares);
3. Construção de um modelo de *machine learning* que classifique as candidaturas em "regulares" ou "laranjas";
4. Teste das hipóteses adotadas durante a construção do modelo;
5. Estimativa da precisão do modelo para detecção de coligações com risco de violarem a cota feminina;
6. Aplicação do modelo às candidaturas de 2020 e criação de uma lista de candidaturas e coligações com risco de serem laranjas e de violarem a cota feminina, respectivamente.

1. Um *proxy* para candidaturas laranjas

Dentre as informações sobre as candidaturas disponibilizadas pelo TSE[2], uma com grande potencial de caracterizar candidaturas laranjas é o número de votos nominais obtidos: dado que tais candidaturas têm como único propósito o cumprimento formal da cota feminina, o número de votos obtidos por elas tende a ser extremamente baixo.

[1] Um dado observável, associado à característica (não observável) que se deseja rastrear.
[2] https://www.tse.jus.br/eleicoes/estatisticas/repositorio-de-dados-eleitorais-1

Para definirmos o número máximo de votos que caracterizaria uma candidatura laranja nas eleições municipais e verificarmos sua qualidade como *proxy*, comparamos a distribuição de número de votos obtidos pelas candidaturas femininas em 2012 e 2016 (depois da implementação da cota de 30%, feita em 2009) com a distribuição em 2004 e 2008.

A Fig. 1 mostra que, depois de 2009, o número de candidaturas femininas com pouquíssimos votos disparou em comparação com candidaturas com mais votos. Além disso, a distribuição de candidaturas masculinas por quantidade de votos pouco se alterou no período[3] (servindo de grupo controle), indicando que a diferença observada nas candidaturas femininas está relacionada à introdução da cota de 30%.

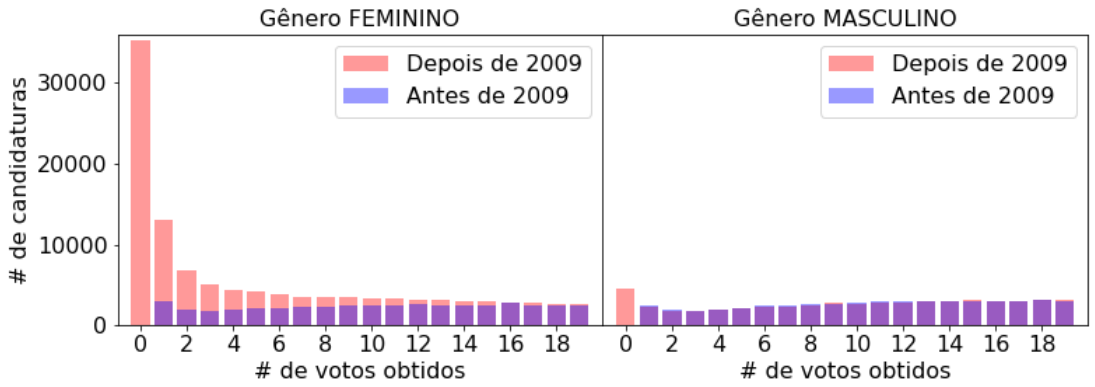

Figura 1: Quantidade de candidaturas femininas (painel esquerdo) e masculinas (painel direito) que obtiveram cada quantidade de votos, antes e depois de 2009. As barras são semi-transparentes e estão sobrepostas. A quantidade de votos anteriores a 2009 foi renormalizada para dar conta do aumento no número total de candidaturas com o tempo.

Assumindo que: (1) candidaturas femininas antes de 2009 são todas regulares; e que (2) a fração das candidaturas regulares que obtém até 1 voto se mantém estável com a introdução da cota, estimamos que **as candidaturas femininas com até 1 voto, depois de 2009, são 94% laranjas**. Assim, adotamos tal quantidade de votos como *proxy* de candidatas laranjas. O problema de previsão de candidaturas laranjas então se traduz na identificação, antes das eleições, de candidaturas femininas que receberão 1 voto ou menos.

[3] É possível notar um ligeiro aumento de candidaturas masculinas com zero votos depois de 2009. Esse aumento tem relação com os raros casos nos quais a quota mínima de 30% não era alcançada por candidaturas masculinas.

2. Construção de *features*

Para caracterizarmos as candidaturas, utilizamos as seguintes bases de dados:

1. Do repositório de dados eleitorais do TSE[4]:
    a. Dados pessoais dos candidatos;
    b. Informações sobre os bens declarados pelos candidatos;
    c. Número de vagas para vereador em cada município;
    d. Perfil do eleitorado por município;
    e. Votações nominais dos candidatos (até 2018).
2. Relação de filiados a partidos políticos, do TSE[5];
3. O IDH 2010 dos municípios brasileiros, obtidos do site da ONU[6]; e
4. Alinhamento ao governo das bancadas dos partidos e dos estados em 2019 (calculados a partir dos dados abertos da câmara[7]).

A partir dessas bases, **criamos 35 *features* para cada candidatura**. Uma comparação da média de cada *feature* calculada para candidaturas regulares e laranjas indica que a incidência de laranjas está, tipicamente, associada a:

1. Municípios pequenos;
2. IDH do município mais baixo;
3. Idade da candidata mais baixa (mais próxima da idade média brasileira, enquanto a idade média na política é maior);
4. Nomes de urna curtos (em geral, há apenas o primeiro nome: não há preocupação com nome);
5. Ausência de nomes de urna diferentes (baixa ocorrência de "Maria da Farmácia", "Samara Corajosa", por exemplo);
6. Data de filiação ao partido mais antigas (filiadas que já estavam disponíveis para serem inscritas como candidatas, sem a necessidade de se filiar);
7. Antes da introdução da quota em 2009, o número de candidatas mulheres no município era mais baixo que a média;

[4] https://www.tse.jus.br/eleicoes/estatisticas/repositorio-de-dados-eleitorais-1
[5] http://www.tse.jus.br/partidos/filiacao-partidaria/relacao-de-filiados
[6] https://www.br.undp.org/content/brazil/pt/home/idh0/rankings/idhm-municipios-2010.html
[7] https://dadosabertos.camara.leg.br/

8. Se já participou de eleições anteriores, a candidata obteve pouquíssimos votos;
9. O número de candidatas do partido representam uma fração maior do total de filiadas (muitas candidatas para poucas filiadas);
10. O partido possui poucas filiadas no município em questão;
11. As mulheres representam uma fração menor dos filiados ao partido naquele município;
12. O eleitorado do município é composto por menos mulheres do que a média;
13. A candidata não apresentou declaração de bens (se não há intenção de ganhar, não há grande preocupação com a burocracia);
14. Quando apresenta declaração de bens, registra um número menor deles;
15. Quando apresenta declaração de bens, o valor total é menor;
16. A escolaridade da candidata é, tipicamente, mais baixa;
17. Ocupação não definida (isso possivelmente caracteriza menor preocupação com a burocracia);
18. Ocupações "Dona de Casa" e "Estudante" (possivelmente de mulheres com vínculos familiares com outros políticos/candidatos) são mais comuns;
19. Ocupações que conferem menos visibilidade (menor incidência de políticos, artistas, professores, sacerdotes);
20. Menor incidência de servidores públicos (estes precisam se afastar do emprego por três meses para poder concorrer);
21. Número de urna dificilmente apresenta todos os dígitos iguais;
22. Número de urna dificilmente é uma sequência (e.g. 12345); e
23. Chapa apresenta pouca (ou nenhuma) folga no cumprimento da cota de 30% de mulheres (cumpre a cota raspando).

Ressaltamos que essa comparação não foi utilizada para classificar candidaturas em laranjas ou regulares (isso foi feito pelo modelo de *machine learning* descrito na seção seguinte). Ela serve apenas para esclarecimento aos seres humanos das possíveis diferenças entre tais candidaturas.

3. Construção de um modelo de *machine learning* de classificação

Para construir um modelo de classificação (aprendizagem supervisionada) de candidaturas femininas ao cargo de vereador em "laranjas" e "regulares", utilizamos os dados das eleições municipais de 2012 e 2016. Juntos, eles totalizam **273.669 candidaturas, sendo que 48.253 delas receberam até 1 voto (correspondendo a 17,6% do total)**.

Os dados foram separados em três subconjuntos disjuntos, denominados: amostra de treinamento; amostra de validação; e amostra de teste. A ideia básica é que o modelo aprende as diferenças entre os dois tipos de candidaturas a partir dos exemplos da amostra de treinamento (com rótulos de "laranja" e "regular"). Depois, para medir sua performance, o modelo faz previsões para a amostra de teste, nunca antes vista por ele (e sem os rótulos), e essas previsões são comparadas com os rótulos. Uma vez que o modelo teve sua estrutura desenhada (i.e. teve seus hiperparâmetros ajustados) e sua performance medida, ele é treinado com todas as 273.669 candidaturas e fica pronto para fazer previsões para 2020[8].

Para evitar que o modelo adotasse estratégias e características pouco estáveis no tempo (que não são reprodutíveis de uma eleição para outra), utilizamos como amostra de treinamento os dados de 2012 e, como amostra de validação e de teste, subconjuntos dos dados de 2016. O modelo selecionado, junto com seus hiperparâmetros e as características utilizadas, foi aquele que maximizou a métrica de desempenho F_1[9] quando testado na amostra de validação (com dados de 2016). **Em outras palavras, o método de classificação escolhido foi aquele que melhor previu candidaturas laranjas em 2016 a partir de exemplos de 2012: um modelo Random Forest**[10]. **Sua precisão (fração das candidaturas classificadas como laranjas que de fato eram laranjas) foi de 63% na amostra de teste. A título de comparação, a classificação aleatória de candidatas como laranjas tem precisão estimada de 15% na mesma amostra.**

É esperado que a qualidade das previsões degrade com o tempo, isto é, que um modelo testado e avaliado em dados de 2016 tenha uma

[8] A amostra de validação é utilizada para fazer testes e ajustes finos ao modelo durante sua construção.
[9] https://en.wikipedia.org/wiki/F-score
[10] https://en.wikipedia.org/wiki/Random_forest

performance um pouco pior em 2020. Para estimar tal nível de degradação, otimizamos um modelo Random Forest para realizar previsões para 2012, e comparamos sua precisão em 2012 com a precisão em 2016. Como observamos uma perda de 5% na precisão, **estabelecemos 58% como precisão esperada para 2020**.

4. Testando as hipóteses adotadas

A construção de um modelo para 2020 a partir dos dados de 2012 e 2016 está baseado na hipótese de que as características das candidaturas (especialmente as que diferenciam laranjas das regulares) não se alteram de maneira significativa ao longo do tempo. Essa hipótese foi analisada das formas descritas abaixo.

Primeiro, verificamos como a distribuição das características dos candidatos evolui com o tempo, de 2012 a 2020 (exemplos são apresentados nas Figs. 2 e 3). É possível notar algumas pequenas mudanças provavelmente ligadas a variações demográficas de longo prazo, como o aumento da escolaridade, aumento da participação feminina e redução da fração de casados. Entretanto, não encontramos diferenças que sugiram uma mudança abrupta e significativa no perfil das candidaturas tal qual a observada com a introdução da quota de 30% em 2009[11].

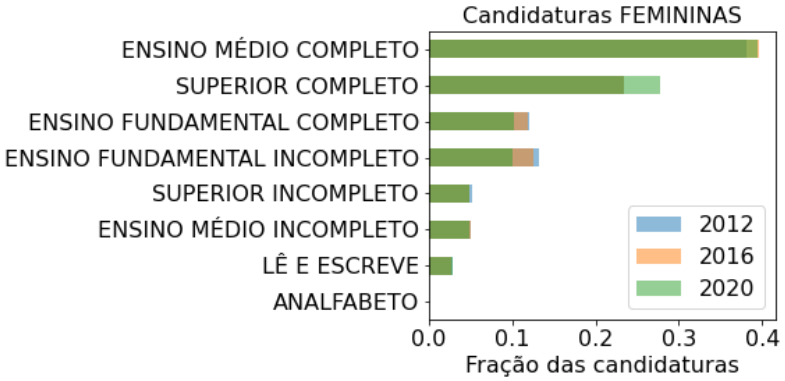

Figura 2: Comparação da distribuição das candidaturas femininas por grau de instrução, para os

[11] O aumento da categoria "outros" para a descrição da ocupação foi a maior variação observada no período entre todas as características analisadas, e pode estar ligada a menor adequação, com o passar do tempo, das categorias definidas pelo TSE.

anos de 2012, 2016 e 2020. As barras são semi-transparentes e estão sobrepostas.

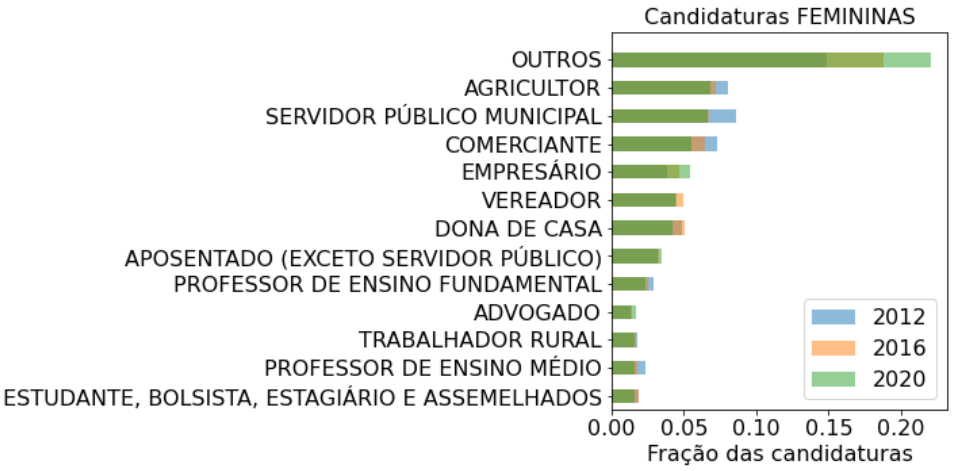

Figura 3: Comparação da distribuição das candidaturas femininas por ocupação (apenas as 13 mais frequentes), para os anos de 2012, 2016 e 2020. As barras são semi-transparentes e estão sobrepostas.

Nós também procuramos por possíveis candidaturas em 2020 que destoassem das de 2016 (ou seja, que fossem *outliers* em relação às candidaturas de 2016). Para tanto:

- treinamos um modelo de *novelty detection* denominado *one-class Support Vector Machine*[12] para encontrar um intervalo nas características[13] que englobasse 99,9% das candidaturas de 2016;
- em seguida, selecionamos as candidaturas de 2016 e de 2020 que não estavam incluídas nesse intervalo (os *outliers*);
- os *outliers* foram agrupados em 15 grupos (*clusters*) de acordo com sua semelhança (utilizando a técnica *k-means*[14]), e a composição de cada grupo (em termos de fração das candidaturas de cada ano) foi calculada;
- eventuais candidaturas de um novo tipo (não presentes em 2016) seriam identificadas por grupos compostos majoritariamente por candidaturas de 2020, sem presença significativa de candidaturas de 2016.

[12] https://dl.acm.org/doi/10.1162/089976601750264965
[13] Em termos técnicos, uma região do espaço de parâmetros.
[14] https://en.wikipedia.org/wiki/K-means_clustering

Essa análise não identificou o surgimento de novos tipos de candidatura em 2020. Uma representação gráfica do agrupamento dos *outliers* é apresentada na Fig. 4, onde vemos que não existem candidaturas de 2020 que destoam das candidaturas de 2016 (i.e. candidaturas de 2020 isoladas de candidaturas de 2016).

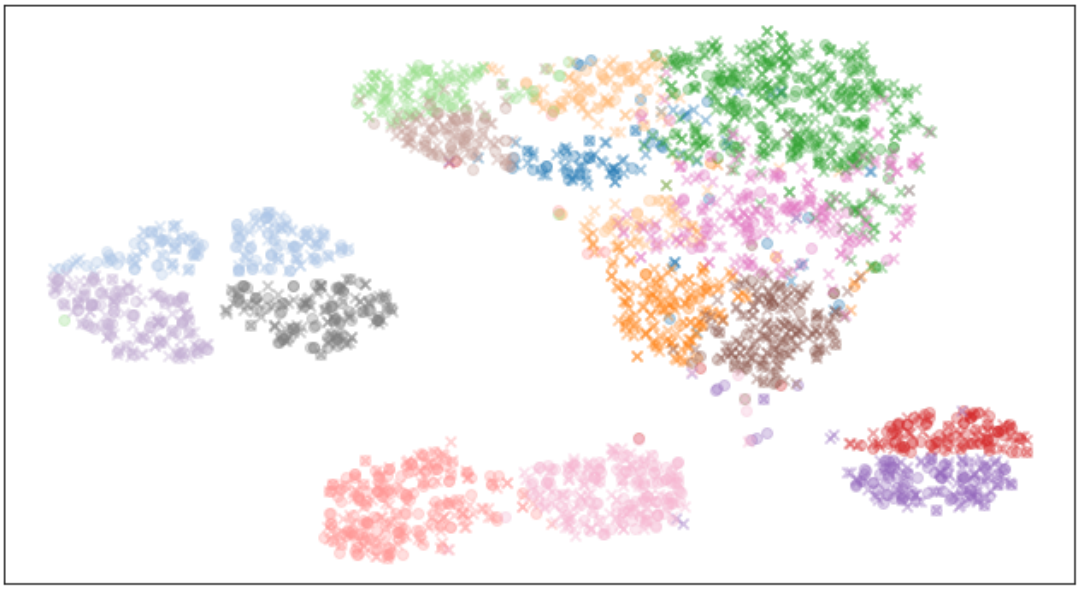

Figura 4: Cada círculo e cruz na figura representa uma candidatura *outlier* de 2016 e de 2020, respectivamente. As candidaturas estão distribuídas na imagem de acordo com sua semelhança (através da técnica t-SNE[15]), e cada cor representa um dos 15 agrupamentos.

Por último, analisamos o impacto da exigência de haver financiamento para campanhas femininas, introduzida em 2018[16]. Um efeito notável dessa regra foi reduzir o número de candidatas mulheres com pouquíssimos votos, conforme mostra a Fig. 5 para eleições de deputados. Essa mudança poderia afetar o resultado das previsões de 2020 caso o perfil das candidatas laranjas também tivesse se alterado significativamente.

[15] https://en.wikipedia.org/wiki/T-distributed_stochastic_neighbor_embedding

[16] https://www.tse.jus.br/imprensa/noticias-tse/2018/Maio/fundo-eleitoral-e-tempo-de-radio-e-tv-devem-reservar-o-minimo-de-30-para-candidaturas-femininas-afirma-tse

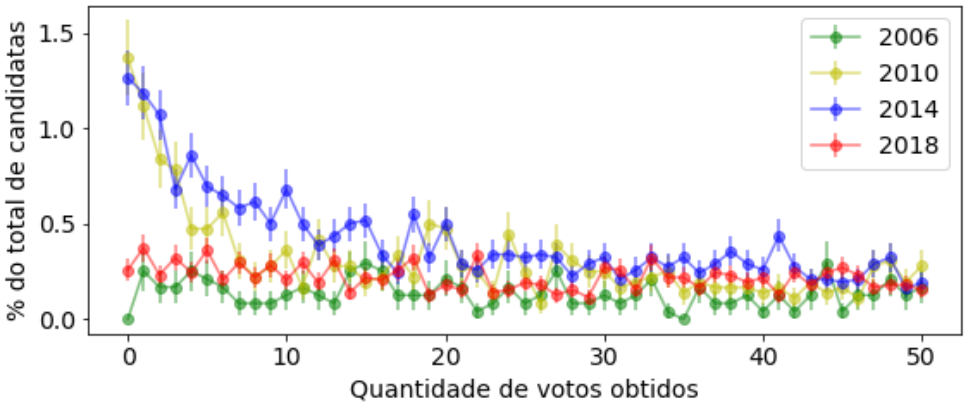

Figura 5: Porcentagem do total de candidatas mulheres a cargos de deputado (estadual, federal e distrital) que recebeu cada quantidade de votos nominais, para as eleições de 2006, 2010, 2014 e 2018.

Para testar essa possibilidade, reproduzimos a construção e teste do modelo descrito na seção 3 para as eleições de deputados. Isto é: utilizamos o ano de 2010 como amostra de treinamento e subconjuntos do ano de 2014 como amostras de validação e de teste.

Devido a menor incidência de candidaturas laranjas nas eleições de deputados (e menor quantidade de exemplos), a precisão do modelo, quando avaliada na amostra teste, ficou em 36%. O ponto crucial é que, quando esse mesmo modelo foi utilizado para prever candidaturas laranjas em 2018, obtivemos uma precisão comparável de 39%[17]. Ou seja: não foi possível observar uma redução na precisão ao aplicar um modelo treinado em candidaturas sem a exigência de financiamento de campanha a candidaturas com tal exigência.

5. Cálculo do risco de coligações violarem a cota feminina

Uma coligação[18] pode violar a cota mínima de 30% de candidaturas femininas nas eleições a vereador de duas formas:

[17] O baixo número de exemplos de candidaturas laranjas entre deputados também amplia a flutuação estatística na medida de precisão. Esta foi estimada em ±10%.

[18] Como nas eleições de 2020 as coligações para eleições proporcionais foram proibidas, o significado de coligação neste documento fica sendo o conjunto de candidaturas de um partido num dado município.

1. Caso menos de 30% das candidaturas aptas a concorrer forem femininas (que chamamos de **violação legal**);
2. Caso a cota mínima de 30% seja alcançada com o uso de candidaturas laranjas.

Na segunda forma de violação, não podemos afirmar que ela esteja de fato ocorrendo, mas podemos selecionar coligações com risco de violar a cota. Isso foi feito utilizando as estimativas de risco das suas candidaturas femininas serem laranjas.

Primeiro, calculamos quantas candidaturas femininas uma coligação possui além do mínimo exigido em lei (que chamamos de excedente). Coligações com excedente negativo, por exemplo, incorreram em violação legal da cota. Em seguida, estimamos a probabilidade da coligação com um certo número de candidaturas femininas ter entre elas, no mínimo, uma quantidade de laranjas igual ao excedente, mais uma[19]. Tipicamente, as coligações em risco possuem excedente nulo, de maneira que basta que uma das candidaturas femininas seja laranja para que a coligação viole a cota.

Aplicando esse método à amostra de teste de 2016, obtivemos uma precisão de 90% ao prever coligações que violam a cota por usarem candidaturas laranjas. Levando em conta uma possível degradação do modelo, estabelecemos como 80% a precisão esperada para 2020.

6. Candidaturas e coligações com risco em 2020

O modelo e o método descritos acima foram aplicados nas candidaturas a vereador de 2020 para estimar, antes da realização das eleições em 15 de novembro, seu risco de serem laranjas (e, a partir destes, estimar o risco da coligação violar a cota feminina).

A tabela abaixo apresenta, nessa ordem: o total de candidaturas a vereador apresentadas nas eleições de 2020 (masculinas e femininas); o total de candidaturas femininas apresentadas; o número de candidaturas femininas consideradas pelo TSE como aptas a concorrer (ou ainda em

[19] Para esse cálculo, utilizamos as probabilidades das candidaturas serem laranjas, calculado pelo algoritmo Random Forest, como probabilidades individuais de ensaios de Bernoulli e calculamos a probabilidade de obtermos um certo número de laranjas através da distribuição binomial de Poisson.

julgamento); a estimativa do número de laranjas esperado em 2020, se a fração observada em 2016 se repetir (entretanto, note que, **entre as eleições de 2014 e 2018, houve uma redução de 20% na fração de laranjas, provavelmente devido à exigência de financiamento das candidaturas femininas**); e o número de candidaturas femininas selecionadas como com risco de serem laranjas (com precisão estimada de 58%).

Eleições proporcionais 2020 Contabilidade de candidaturas	
Total de candidaturas	520.571
Candidaturas femininas	180.578
Candidaturas femininas aptas	170.820
Estimativa de laranjas femininas[20]	26.725
Selecionadas como em risco	2.558

Em comparação às candidatas que tipicamente obtém mais votos, o perfil das candidaturas laranjas tende a ser menos branca, menos escolarizada, mais jovem e residir em municípios mais pobres. Paralelamente, modelos de *machine learning* tendem a generalizar o rótulo (nesse caso, de laranja) para exemplos semelhantes (nesse caso, candidaturas com o perfil mencionado acima). **Isso introduz um viés no modelo, que acaba por selecionar, com frequência excessiva, esse perfil mencionado.** A Fig. 6, por exemplo, mostra que candidatas reais negras tem quase 3 vezes mais chance de serem erroneamente apontadas como laranjas do que candidatas reais brancas; e isso mesmo com o modelo não fazendo uso da informação sobre raça de maneira direta.

[20] Com base em 2016. Note, entretanto, que a exigência de financiamento das campanhas femininas, instituída em 2018, pode reduzir o número de laranjas para 20% desse valor, tomando por base as eleições de 2014 e 2018.

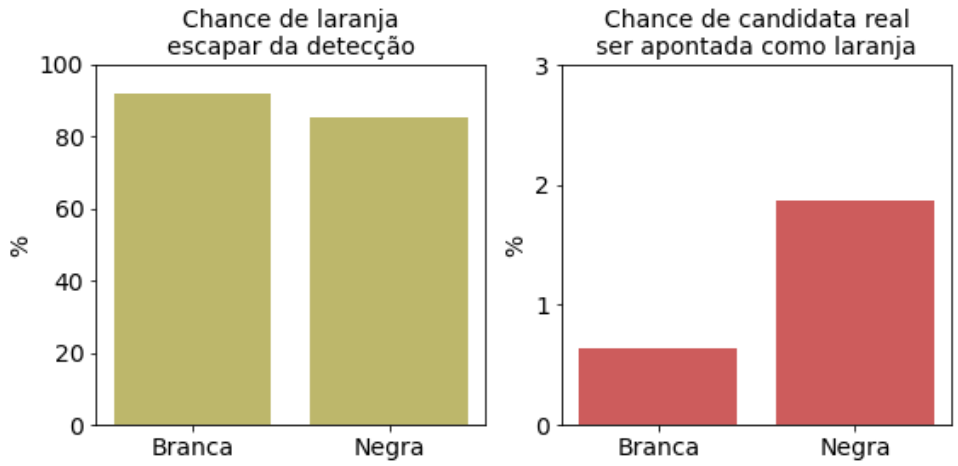

Figura 6: Probabilidades do modelo cometer erros na classificação de candidaturas por cor/raça da candidata, calculadas com dados de 2016. O painel esquerdo mostra a fração das candidatas laranjas (i.e. que obtiveram 0 ou 1 votos) que não foram selecionadas pelo modelo como possíveis laranjas. O painel direito mostra a fração das candidatas reais (i.e. que obtiveram mais do que 1 voto) que são erroneamente identificadas pelo modelo como laranjas.

A fim de mitigar esse efeito - que poderia prejudicar pessoas tipicamente mais vulneráveis e que não são laranjas -, selecionamos uma amostra aleatória (mas com pesos) de 50 candidaturas dentre as 2.558 identificadas como em risco de maneira a tornar a distribuição de raça, estado, idade e classe mais próximas do observado para as candidaturas como um todo.

Por fim, a tabela abaixo apresenta a contagem de coligações das eleições proporcionais de 2020 de acordo com os seguintes critérios (nessa ordem): total de coligações apresentadas; total de coligações que, em 30/11, ainda apresentavam ao menos uma candidatura apta ou em julgamento; estimativa, com base nos dados de 2016, do número de coligações que violaram a cota, seja legalmente ou por utilizarem candidaturas laranjas; o número de coligações de 2020 selecionadas pelo modelo como com risco de violar a cota feminina (seja a exigência formal, seja através do uso de laranjas); dentre essas coligações, aquelas que violaram legalmente a cota (não possuem candidatas aptas em número suficiente); e número de coligações com risco de violarem a cota por fazerem uso de candidaturas laranjas.

Eleições proporcionais 2020 Contabilidade de coligações	
Total de coligações	40.799
Coligações com cand. aptos	40.320
Estimativa de violações da cota	15.471
Coligações selecionadas	931
Selecionadas com violação legal	737
Selecionadas com risco de laranja	194

Principais resultados do estudo

Os principais resultados do estudo são previsões para as eleições de 2020:

1. Selecionamos uma lista de "risco laranja": 2.558 candidatas que apresentam 58% de chance de serem laranjas;
2. Da lista acima, selecionamos uma sub-amostra de 50 candidaturas corrigindo o viés do algoritmo que tendia a destacar mais pessoas negras, jovens e de certos estados;
3. Selecionamos uma lista de 194 chapas que apresentam 80% de chance de fazerem uso de candidaturas laranjas;
4. Encontramos 737 chapas que não atingiram o mínimo de 30% de candidaturas femininas aptas a concorrer;
5. Estimamos que o número total de laranjas inscritas como candidatas (selecionadas ou não pelo nosso modelo) está entre 5 e 27 mil.

ANEXO - Avaliação pós-eleições

Após a realização do primeiro turno das eleições de 2020, no dia 15 de novembro, nós baixamos seus resultados (as votações nominais) e fizemos uma avaliação do desempenho do modelo.

Grande queda no número de candidatas com 0 ou 1 votos

Em 2016, 15% das candidatas mulheres obtiveram 0 ou 1 votos, totalizando 21.262 pessoas. Uma comparação com eleições anteriores a 2009, ano de implementação da cota de gênero, indica que 94% delas seriam laranjas. Se a porcentagem de 15% se mantivesse em 2020, teríamos cerca de 25.600 mulheres com essa quantidade de votos. Nesse caso, porque restringimos nossa previsão para um mínimo de laranjas de 5 mil?

Fizemos isso porque foi observada uma queda significativa no número de candidatas com 0 ou 1 votos entre as eleições de 2014 e 2018 (veja a Fig. 5). Essa queda abrupta poderia ter relação com a exigência de financiamento de candidaturas femininas em 2018[21] ou com outras mudanças legais que tornam mais grave o uso de candidaturas laranjas[22]. Por esse motivo, consideramos possível que uma queda semelhante acontecesse entre as candidatas ao cargo de vereador. Como mostra a Fig. 7, isso realmente ocorreu: **o número de candidatas com 0 ou 1 votos em 2020 foi de 6 mil, próximo (e acima) do mínimo previsto**.

[21] https://www.camara.leg.br/tv/541668-nova-regra-para-as-candidaturas-femininas/

[22] Em 2019, por exemplo, o TSE considerou que a burla nas cotas para as mulheres configura uma fraude passível de cassação de toda a chapa que tenha se beneficiado da fraude: https://radios.ebc.com.br/revista-rio/2020/10/eleicoes-terao-regras-mais-rigidas-para-garantir-candidaturas-femininas

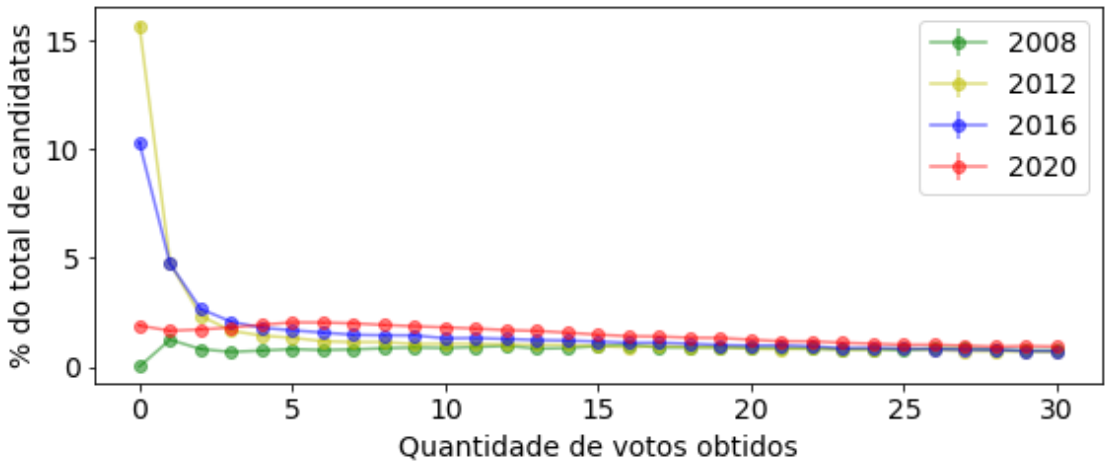

Figura 7: Porcentagem do total de candidatas mulheres a cargos de vereador que recebeu cada quantidade de votos nominais, para as eleições de 2008, 2012, 2016 e 2020.

Maior dificuldade de validação do modelo

Até 2018, a quantidade de candidatas que obtinham 0 ou 1 votos era relativamente grande, e sua grande maioria era composta por laranjas. Isso nos permitia testar a capacidade do modelo em identificar laranjas verificando a porcentagem das candidatas selecionadas que obtiveram 0 ou 1 votos. Nós certificamos que 63% das candidatas de 2016 selecionadas pelo modelo como laranjas possuíam essa quantidade de votos.

A brutal queda no número de candidatas com 0 ou 1 votos - que pode ser causada por uma redução do número de laranjas, pelo fato de laranjas passarem a obter mais votos, ou por uma combinação desses dois motivos - fez com que, das candidatas de 2020 selecionadas pelo modelo, 21% tivessem obtido no máximo 1 voto. Essa aparente perda de precisão (de 63% para 21%) ocorreu por dois motivos:

- A seleção dessas candidatas se tornou "procurar uma agulha no palheiro": enquanto, em 2016, a chance de se encontrar tal candidata por acaso era de 15%, agora ela é de 3,5%. Mesmo que o modelo erre na mesma taxa que a observada em dados de 2016, agora, para cada candidata laranja, existem uma quantidade maior de candidatas regulares similares a laranjas para confundi-lo.

- Candidatas laranjas corretamente selecionadas pelo modelo mas que obtiveram mais do que um voto contam como um "erro" se usarmos a quantidade de votos para definir laranjas.

Antes de apresentar evidências desses mecanismos, vale reforçar que, **mesmo que o modelo seja avaliado pela capacidade de seleção de pessoas com 0 ou 1 votos, ele ainda apresenta um bom desempenho: em comparação com uma busca cega por tais candidatas, o modelo aumenta a probabilidade de acerto de 3,5% para 21%, um aumento de 6 vezes.** A Fig. 8 compara histogramas (distribuição de candidatas pelo número de votos obtidos) das selecionadas e ignoradas pelo modelo. Podemos perceber que **64% das candidatas selecionadas como laranjas pelo modelo obtiveram no máximo 10 votos, enquanto que tal fração é de 20% para as candidatas não selecionadas**.

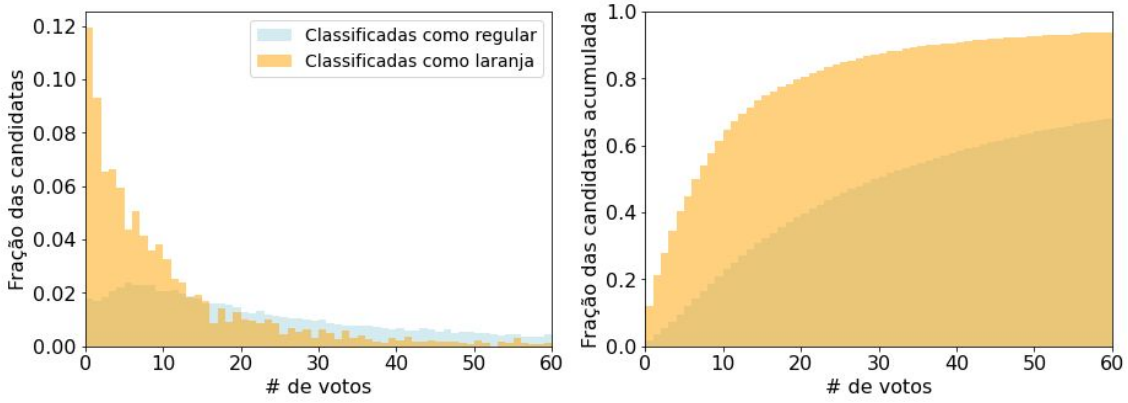

Figura 8: O painel esquerdo mostra a fração das candidatas que obtiveram cada quantidade de votos (histograma normalizado) para aquelas selecionadas pelo modelo como laranjas e para as ignoradas pelo modelo (i.e. classificadas como regular). As barras de cada histograma são semi-transparentes e estão sobrepostas. O painel da direita mostra a fração de candidatas acumulada até um certo número de votos, isto é, mostra a fração de candidatas que obtiveram um certo número de votos ou menos.

Menos laranjas e laranjas com mais votos

Os dois efeitos mencionados na seção anterior - a redução do número de laranjas e o aumento do número de laranjas com mais de 1 voto em 2020 - foram simulados em dados de 2016: nós retiramos da amostra algumas candidatas com 0 ou 1 votos e para outras nós

aumentamos seu número de votos obtidos (todas selecionadas aleatoriamente). Esses dois números precisam se equilibrar de maneira que a fração de candidatas com 0 ou 1 votos seja de 3,5%, a observada em 2020.

A Fig. 9 mostra como, dentre as candidatas selecionadas pelo modelo, a fração de candidatas com 0 ou 1 votos varia em função dessas modificações aplicadas aos dados de 2016. Os dados de 2020 poderiam ser explicados com uma redução de 55% no número de laranjas em relação a 2016 e com 25% das candidatas laranjas recebendo mais de 1 voto, aproximadamente.

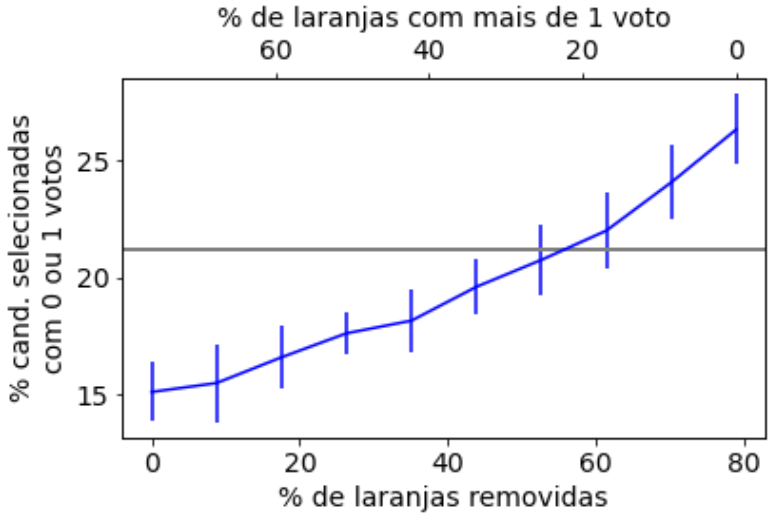

Figura 9: Como a precisão do modelo (fração de candidatas selecionadas que obtiveram 0 ou 1 votos) depende da remoção de laranjas e do aumento do número de laranjas com mais de 1 voto, estimado a partir de simulações sobre dados de 2016. Essas duas modificações dos dados precisam ser tais que a fração de candidatas com 0 ou 1 votos (independentemente de seleção pelo modelo) seja de 3,5% do total. A linha horizontal cinza mostra a precisão medida em 2020.

Outro fato que corrobora que laranjas podem receber mais do que 1 voto e que a capacidade do modelo em identificar laranjas é maior do que sua capacidade de identificar pessoas com 0 ou 1 votos foi a investigação feita pelo jornal *O Estado de São Paulo*[23]. **Eles conseguiram entrar em contato com três candidatas selecionadas pelo modelo, e todas**

23 https://politica.estadao.com.br/noticias/eleicoes,estudo-indica-ao-menos-5-mil-candidatas-laranjas-nas-eleicoes-2020,70003512533

confirmaram ser laranjas, apesar de apenas uma ter recebido 1 voto (as duas outras receberam 3). Se a probabilidade do modelo selecionar laranjas fosse de apenas 21%, a chance de encontrar 3 laranjas em 3 tentativas seria de 0.9%, um número excessivamente baixo. Ou seja: 21% é a probabilidade do modelo de encontrar candidatas com 0 ou 1 votos, sendo que a probabilidade dele encontrar laranjas é bem maior. Supondo uma probabilidade mais realista de se obter 3 sucessos em 3 tentativas - por exemplo, de 5% -, a probabilidade do modelo de acertar ao selecionar uma laranja teria que ser de, no mínimo, 37%.

Resumo e conclusões

Em 2020 observou-se, em comparação com 2012 e 2016, uma queda significativa no número de candidatas com 0 ou 1 votos. Isso pode ser explicado pela combinação de uma redução no número de laranjas e do aumento do número de laranjas com mais de 1 voto a partir de 2018. Como nosso modelo foi treinado com dados de 2012 e 2016, época em que 94% das candidatas com esses votos eram laranjas, ele aprendeu a identificar laranjas. Isso o permite identificar laranjas mesmo quando elas recebem mais de um voto.

A chance de acerto do modelo ao apontar candidatas em 2020 que receberão 0 ou 1 votos é fácil de estimar: 21%. Esse número é 6 vezes maior do que a probabilidade de encontrar tais candidatas no chute, de 3,5%. Por sua vez, estimar a probabilidade de acerto do modelo ao apontar candidatas laranjas em 2020 é mais difícil de ser feita: isso exigiria uma investigação como a feita pelo Estadão, mas com um grande número de candidatas. Com base nos resultados da matéria do Estadão e nas análises feitas com dados de 2016, estimamos que tal precisão esteja entre 37% e 63%.

Gabinete Compartilhado

Coordenação

Chefe de Gabinete
José Frederico

Cientista de dados
Henrique Xavier

Revisão
Ana Marina de Castro