Universidade Federal da Paraíba

Centro de Informática

Programa de Pós-Graduação em Informática

Um modelo para a detecção das mudanças de posicionamento dos deputados federais

Vitor Márcio Paiva de Sousa Baptista

Dissertação submetida à Coordenação do Curso de Pós-Graduação em Informática da Universidade Federal da Paraíba como parte dos requisitos necessários para obtenção do grau de Mestre em Informática.

Área de Concentração: Ciência da Computação
Linha de Pesquisa: Computação Distribuída | Sinais, Sistemas Digitais e Gráficos

Alexandre Nóbrega Duarte
(Orientador)

João Pessoa, Paraíba, Brasil

Resumo

No Brasil, existem ferramentas para o acompanhamento do comportamento dos parlamentares em votações nominais, tais como o Basômetro do jornal O Estado de São Paulo e o Radar Parlamentar. Essas ferramentas são usadas para análises tanto por jornalistas, quanto por cientistas políticos. Apesar de serem ótimas ferramentas de análise, sua utilidade para monitoramento é limitada por exigir um acompanhamento manual, o que se torna muito trabalhoso quando consideramos o volume de dados. Somente na Câmara dos Deputados, 513 parlamentares participam em média de mais de 400 votações nominais por legislatura. É possível diminuir a quantidade de dados analisando os partidos como um todo, mas em contrapartida perdemos a capacidade de detectar movimentações de indivíduos ou grupos intrapartidários como as bancadas. Para diminuir esse problema, desenvolvi neste trabalho um modelo estatístico que detecta quando um parlamentar muda de posicionamento, entrando ou saindo da coalizão governamental, através de estimativas de pontos ideais usando o W-NOMINATE. Ele pode ser usado individualmente ou integrado a ferramentas como o Basômetro, oferecendo um filtro para os pesquisadores encontrarem os parlamentares que mudaram mais significativamente de comportamento. O universo de estudo é composto pelos parlamentares da Câmara dos Deputados no período da 50ª até a 54ª legislaturas, iniciando no primeiro mandato de Fernando Henrique Cardoso em 1995 até o início do segundo mandato de Dilma Rousseff em 2015.

Palavras-chave: Análise legislativa, Ciência política, Ciência de dados, Modelos preditivos, Aprendizagem de máquina.

Abstract

In Brazil, there are tools for monitoring the behaviour of legislators in rollcalls, such as O Estado de São Paulo's Basômetro and Radar Parlamentar. These tools are used both by journalists and political scientists for analysis. Although they are great analysis tools, their usefulness for monitoring is limited because they require a manual follow-up, which makes it a lot of work when we consider the volume of data. Only in the Chamber of Deputies, 513 legislators participate on average over than 400 rollcalls by legislature. It is possible to decrease the amount of data analyzing the parties as a whole, but in contrast we lose the ability to detect individuals' drives or intra-party groups such as factions. In order to mitigate this problem, I developed a statistical model that detects when a legislator changes his or her position, joining or leaving the governmental coalition, through ideal points estimates using the W-NOMINATE. It can be used individually or integrated to tools such as Basômetro, providing a filter for researchers find the deputies who changed their behaviour most significantly. The universe of study is composed of legislators from the Chamber of Deputies from the 50th to the 54th legislatures, starting in the first term of Fernando Henrique Cardoso in 1995 until the beginning of the second term of Dilma Rousseff in 2015.

Keywords: Legislative Analysis, Political Science, Data Science, Predictive Models, Machine Learning.

Agradecimentos

Devo todas as conquistas de minha vida à minha família. Foi o seu trabalho, amor e dedicação que me ensinaram e me permitiram fazer o que fiz. Se todos chegamos aonde estamos por nos apoiarmos nos ombros de gigantes, foram eles e elas os primeiros gigantes nos quais me apoiei. Em especial, agradeço à minha mãe, Dione, e meus avós maternos, Dalvinha e Paulo.

Agradeço à minha esposa, melhor amiga e coorientadora não-oficial Samara. Sem sua ajuda, esse trabalho não seria possível e minha vida seria muito mais solitária.

Agradeço também ao meu orientador, Alexandre. Só o conheci ao me inscrever no mestrado, mas ao longo desses anos tive certeza de que não poderia ter tido mais sorte nessa escolha. Por sua orientação técnica, mas principalmente pelo seu interesse indiscutível nessa área de pesquisa. Em outra situação, acredito que ele mesmo teria escrito esse trabalho, o que é prova inegável da nossa sintonia.

Agradeço às professoras Andréa Freitas e Thaís Gaudêncio, que aceitaram participar da minha banca, emprestando seu tempo e conhecimento para a melhoria deste trabalho.

Agradeço aos pesquisadores e funcionários do Centro Brasileiro de Análise e Pesquisa (CEBRAP), cuja contribuição à área da Ciência Política é incalculável, no Brasil e no mundo. Em especial, gostaria de agradecer a Andréa Freitas, ao Samuel Moura e ao Maurício Izumi, que me ajudaram muito a validar as ideias que discuti nesse trabalho, e ao Paulo Hubert, que me auxiliou a acessar o banco de dados legislativos do CEBRAP, do qual extraí a lista usada de coalizões. Além, é claro, à Argelina Figueiredo e ao Fernando Limongi, cuja pesquisa foi um divisor de águas no pensamento da Ciência Política brasileira.

Ao longo do tempo, o contato com pessoas interessadas na intersecção entre computação, política e jornalismo foi abrindo meus olhos para essa nova área que acho extremamente importante. Isso foi possibilitado, principalmente, pela criação do grupo Transparência Hacker por, entre tantos outros, Pedro Markun e Daniela Silva. Através desse grupo, conheci pessoas fenomenais como os jornalistas Daniel Bramatti, José Roberto de Toledo e Amanda

Rossi que, junto com o Diego Rabatone, formavam o Estadão Dados, onde tive o prazer de trabalhar por uma semana durante o segundo turno das eleições de 2012.

Agradeço aos amigos criados durante a organização do Encontro de Software Livre da Paraíba (ENSOL), em especial a Rodrigo Vieira e Anahuac de Paula Gil, os principais responsáveis no meu amadurecimento com relação a software e cultura livres.

Agradeço também aos amigos gaúchos Leonardo Tartari e Thiago Bueno, companheiros de vários hackathons. Foram eles que aprofundaram em mim o interesse pela visualização de dados, que foi uma das razões que me fizeram entrar na Open Knowledge Foundation (OKF).

A OKF é uma ONG inglesa que trabalha com dados abertos. Durante os anos que trabalhei nela, tive oportunidade de conhecer diversas pessoas que me ajudaram a me aprofundar nessa área. Agradeço em especial ao time do CKAN e aos fundadores da Open Knowledge Foundation Brasil.

Por último, mas de forma alguma menos importante, agradeço aos amigos brutais, os irmãos e irmãs que encontrei durante a vida. Em especial ao Pedro Guimarães que, além de amigo e parceiro em diversos projetos, acabou se tornando meu cunhado.

A todas essas pessoas e muitas outras, dedico este trabalho.

“Geeks like to think that they can ignore politics. You can leave politics alone, but politics won’t leave you alone.”

— Richard Stallman

Conteúdo

Lista de Siglas

ACP Análise de Componentes Principais

API *Application Programming Interface*

AUC área sob a curva ROC

CEBRAP Centro Brasileiro de Análise e Pesquisa

ELLO Estudos Legislativos e Análise Política

GBM *Stochastic Gradient Boosting*

GLM Modelo Linear Generalizado

IR índice de Rice

MCA *Multiple Correspondence Analysis*

NNET rede neural

PSOL Partido Socialismo e Liberdade

PT Partido dos Trabalhadores

RF *Random Forest*

ROC *Receiver Operating Characteristic*

SVM *Support Vector Machine*

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Neste capítulo, serão descritos o que motivou o desenvolvimento deste trabalho (Seção 1.1), juntamente com a definição do problema e objetivos da pesquisa, a metodologia seguida, publicações relacionadas e, por fim, será resumida a estrutura do restante da dissertação.

1.1 Motivação

O acompanhamento das atividades dos legisladores é extremamente importante, pois são eles que alteram as regras do jogo para a vida no Brasil, afetando todos que têm alguma relação com o Brasil, estando ou não em solo brasileiro. As grandes empresas, com recurso para investir, reconhecem a importância de acompanhar de perto a atividade legislativa, seja passiva ou ativamente através de *lobbying*. Infelizmente, com exceção das leis que são divulgadas na mídia, como aumentos do salário mínimo ou, recentemente, a redução da maioridade penal, a maioria dos cidadãos não se interessa por essa área, seja por falta de tempo, conhecimento ou simplesmente falta de interesse.

Jornalistas políticos exercem um papel fundamental nesse sentido, traduzindo os termos técnicos e jurídicos usados pelos parlamentares em uma forma que possa ser mais facilmente compreendida pelo cidadão comum. Entretanto, como essa análise demanda bastante tempo, ela acaba se restringindo aos temas mais polêmicos, que atingem um maior número de pessoas. Esses temas são muito importantes, mas não suficientes: se eu trabalho numa ONG de

preservação do meio ambiente, por exemplo, meu maior interesse é em projetos de lei que versem sobre essa área.

Percebendo essa necessidade, empresas como o Estudos Legislativos e Análise Política (ELLO) no Brasil e a FiscalNote nos Estados Unidos, criaram ferramentas que facilitam esse monitoramento personalizado. Seus produtos permitem que o usuário defina suas áreas de interesse (por exemplo, meio ambiente ou mobilidade urbana), recebendo resumos periódicos do que as afeta, inclusive com previsões da probabilidade de aprovação dos projetos de lei relacionados a elas. Apesar disso, por serem ferramentas pagas, seu uso ainda é restrito.

Um dos fatores mais importantes no comportamento dos parlamentares é o conflito governo/oposição (LEONI, 2002; DESPOSATO, 2005; FREITAS; IZUMI; MEDEIROS, 2012; IZUMI, 2013). Diante disso, quando um parlamentar muda de lado, seja migrando de partido ou quando seu próprio partido se une (ou deixa) à coalizão governamental, é de se esperar que seu comportamento também mude. Como, no geral, os partidos são capazes de disciplinar seus filiados a votarem de certa forma, essa mudança de forças pode definir a aprovação ou não de um projeto de lei. Assim, é essencial monitorar essas mudanças para entender as chances que um projeto tem de ser aprovado.

Entretanto, o grande volume de dados gera um desafio. Considerando somente o nível federal, 513 deputados federais e 81 senadores participam de centenas de votações em cada legislatura, tornando difícil o entendimento dos seus padrões de votação. Algumas ferramentas, como o Basômetro e o Radar Parlamentar, foram criadas para tentar diminuir esse problema (ESTADÃO, 2012; TRENTO; LEITE, 2013).

Elas são ferramentas gratuitas que permitem o acompanhamento da taxa de governismo (no caso do Basômetro) ou das posições relativas dos parlamentares (no caso do Radar Parlamentar), auxiliando o cidadão comum a se aproximar do processo legislativo (DANTAS; TOLEDO; TEIXEIRA, 2014). Entretanto, mesmo facilitando bastante, não é fácil analisar um gráfico com 513 pontos (no caso da Câmara) e visualizar o que está mudando ou não. Para diminuir esse desafio, as análises acabam sendo feitas baseadas no comportamento agregado dos partidos, e não dos parlamentares individualmente.

Os partidos são capazes de influenciar o comportamento dos seus parlamentares, man-

tendo taxas de disciplina, em sua maioria, acima de 75% (FIGUEIREDO; LIMONGI, 2001; CHEIBUB; FIGUEIREDO; LIMONGI, 2009; ZUCCO, 2009). Assim, a análise a nível de partido é uma forma importantíssima de entender o comportamento parlamentar. Apesar disso, algumas informações são perdidas ao fazer essa agregação. Para dar um exemplo concreto, o Partido dos Trabalhadores (PT) é, historicamente, um dos partidos brasileiros mais disciplinados. Em 2003, o primeiro ano do primeiro governo Lula, o PT manteve uma disciplina altíssima (98,36 % na Câmara). Apesar disso, três deputados e uma senadora constantemente votavam contrários a indicação do partido, o que acabou resultando na sua expulsão no final do mesmo ano (BREVE, 2003).

Esses parlamentares são os deputados Babá, Luciana Genro e João Fontes e a senadora Heloísa Helena. Após serem expulsos, fundaram um novo partido: o Partido Socialismo e Liberdade (PSOL). Essa movimentação passaria desapercebida ao analisar o PT como um todo, já que só 4 dos quase 100 deputados e senadores do partido tiveram esse comportamento.

O objetivo deste trabalho é o desenvolvimento de um modelo estatístico que determine a chance de um parlamentar ter mudado de posicionamento com base no seu padrão de voto. Esse modelo foi treinado com os dados da 50ª até a 54ª legislaturas, compreendendo o período de 20 anos de 1995, no início do primeiro governo de Fernando Henrique Cardoso, até o início de 2015, no início do segundo governo de Dilma Rousseff.

Com isso, espero criar uma ferramenta para que os cidadãos, jornalistas e cientistas políticos consigam filtrar e ordenar os parlamentares pela intensidade da sua mudança de comportamento. Dessa forma, eles otimizariam o uso do seu tempo, focando em quem está mudando. Esse modelo poderá ser usado separadamente, ou integrado em ferramentas já existentes, como o Basômetro ou o Radar Parlamentar, visando aumentar sua utilidade como ferramentas de monitoramento legislativo.

Essa ferramenta poderá ser útil também para os próprios parlamentares, especialmente os líderes, permitindo que monitorem se seus liderados estão mudando de lado.

Como preditores, o modelo usa os pontos ideais dos parlamentares, estimados pelo algoritmo W-NOMINATE (POOLE; ROSENTHAL, 1985; POOLE, 2005). Além desses pontos,

também replicamos parte da pesquisa de Freitas (2008), que analisa os aspectos temporais das migrações partidárias no Brasil, focando nas mudanças de posicionamento, sejam a partir da migração para partidos de posicionamento oposto (indo do governo para oposição ou vice-versa), ou na entrada ou saída do próprio partido na coalizão governamental.

Uma das principais contribuições deste trabalho é democratizar o acesso a técnicas antes só disponíveis em sistemas pagos de empresas como o ELLO e a FiscalNote.

1.2 Problema de pesquisa

Os partidos e as coalizões governamentais são capazes de influenciar o comportamento dos parlamentares (FIGUEIREDO; LIMONGI, 2001; SANTOS, 2003). Partindo disso, Izumi (2013) mostrou que o comportamento dos senadores muda ao entrar ou sair da coalizão, mas não sabemos se ele muda antes ou depois da oficialização dessa mudança. A pergunta a que buscamos responder neste trabalho é:

É possível detectar a mudança de posicionamento de um deputado federal, com ele entrando ou saindo da coalizão governamental, a partir de uma mudança no seu padrão de votação?

A premissa básica deste trabalho é que, além de ser possível detectar mudanças de posicionamento, essa detecção ocorra antes da oficialização da mudança. Em outras palavras, que o modelo seja capaz de detectar uma mudança de posicionamento antes que ela seja do conhecimento público. Para isto, precisaremos também responder à pergunta:

Os deputados federais mudam seu padrão de votação antes de mudarem de posicionamento?

Alguns autores mostraram que os parlamentares mudam de comportamento ao mudarem de posicionamento (ver Capítulo 3). Apesar disso, não foram encontrados trabalhos que discorram sobre se tal mudança de comportamento ocorre antes ou depois da efetiva mudança de posicionamento. Assim, responder a essa pergunta é uma outra contribuição deste trabalho.

1.3 Objetivos

O objetivo geral desta dissertação é o desenvolvimento e validação de um modelo capaz de determinar a chance de um deputado federal ter mudado de posicionamento em um determinado período.

Para alcançar esse objetivo geral, foram definidos os seguintes objetivos específicos:

- Determinar um conjunto de características a partir das quais seja possível determinar a chance de um parlamentar mudar de posicionamento;
- Descobrir se os parlamentares mudam de comportamento antes de mudarem de posicionamento;
- Analisar diversos modelos estatísticos, buscando qual tem melhor performance na detecção da mudança de posicionamento dos deputados federais brasileiros.

1.4 Metodologia

Para o desenvolvimento desta pesquisa, foram seguidos os seguintes passos:

1. Levantamento bibliográfico sobre análise do comportamento parlamentar, análise da mudança do comportamento parlamentar, e métodos de aprendizado de máquina usados no âmbito da Ciência Política;
2. Extração dos dados de votos e votações a partir da página da Câmara dos Deputados, e da listagem de coalizões a partir do banco de dados legislativo do CEBRAP;
3. Definição da forma para representação desses dados, usando a teoria espacial do voto;
4. Análise dos padrões gerais e temporais dos pontos ideais estimados;
5. Definição de variáveis independentes capazes de serem usadas para diferenciar parlamentares que mudaram de posicionamento dos que não mudaram;

6. Análise de diversos modelos preditivos buscando o que obtêm a melhor performance nesses dados;

7. Validação do modelo final baseado na análise feita na etapa anterior.

1.5 Publicações Relacionadas

Resultados iniciais desta pesquisa foram publicados no artigo "Uma ferramenta para analisar mudanças na coesão entre parlamentares em votações nominais", apresentado no III Brazilian Workshop on Social Network Analysis and Mining (BRASNAM), ocorrido em 2014 (BAPTISTA et al., 2014).

1.6 Estrutura da Dissertação

Esta dissertação é dividida em cinco capítulos, incluindo este introdutório, que apresentou a motivação, problema de pesquisa e objetivos deste trabalho.

No Capítulo 2, Fundamentação Teórica, serão brevemente apresentados os conceitos básicos necessários para entendimento do restante do trabalho. Na Seção 2.1, falarei sobre a Ciência de Dados, com foco no desenvolvimento de modelos preditivos para predição de variáveis categóricas e sua validação através de matrizes de confusão e ferramentas como a curva *Receiver Operating Characteristic* (ROC).

Na Seção 2.2, apresentarei técnicas baseadas na teoria espacial do voto, que permitem colocar um conjunto de parlamentares em um plano cartesiano, com suas posições definidas a partir de seus votos em um conjunto de votações. Existem diversas técnicas para definir essas posições, mas neste trabalho focarei no W-NOMINATE, uma das mais usadas.

Na Seção 2.2.1, será apresentado o problema em comparar pontos ideais ao longo do tempo, descrevendo técnicas que usam "pontes" para diferenciar mudanças causadas por diferenças na agenda legislativa dos períodos das causadas pela mudança de comportamento do parlamentar.

No Capítulo 3, Trabalhos Relacionados, serão resumidos alguns trabalhos, encontrados durante a revisão bibliográfica feita nesta pesquisa, que versam sobre a análise do comportamento parlamentar, a mudança de comportamento e o uso de técnicas de aprendizagem de máquina na Ciência Política.

Explicadas, até este momento, as ferramentas usadas no trabalho e a literatura da área, no Capítulo 4, Desenvolvimento do modelo preditivo, será descrito o processo de criação do modelo, partindo da definição do universo de estudo, coleta e preparação dos dados, gerando estimativas dos pontos ideais dos parlamentares, passando pela análise das características gerais e temporais dos dados para, finalmente, descrever o desenvolvimento e validação do modelo na Seção 4.4, Modelagem.

Por fim, o Capítulo 5, Conclusão, apresenta as conclusões da pesquisa, incluindo suas limitações e possíveis trabalhos futuros.

Capítulo 2

Fundamentação Teórica

Neste capítulo, serão abordados os conceitos necessários para compreensão do restante do trabalho. Na Seção 2.1, explicarei o que é ciência de dados, descrevendo também o processo de desenvolvimento e avaliação de modelos preditivos, com foco em modelos de classificação. Na Seção 2.2, falarei brevemente sobre a teoria espacial do voto e a técnica W-NOMINATE para estimativa de pontos ideais.

2.1 Ciência de Dados

Segundo Stanton (2012), o termo Ciência de Dados (do inglês *Data Science*) é usado para definir uma área emergente que se ocupa da coleta, preparo, análise, visualização, gestão e preservação de conjuntos de dados com o objetivo de extrair conhecimento. Ela usa a ciência da computação como ferramenta para extrair modelos estatísticos a partir de dados relativos a uma área fim, como a ciência política.

Por ser uma área interdisciplinar e relativamente recente (especialmente no Brasil), a distinção do que é ciência de dados e não estatística, matemática ou computação pode ainda não ser muito clara (PORTO; ZIVIANI, 2014). Para facilitar essa diferenciação, Conway (2013) criou o diagrama de Venn da Figura 2.1, que mostra onde se situa a ciência de dados em relação às outras áreas. Nela, vemos que a ciência de dados está na intersecção entre computação, estatística e uma área fim (por exemplo, ciência política). A pesquisa tradicio-

nal, segundo ele, estaria na intersecção da estatística com a área fim, enquanto a pesquisa de aprendizagem de máquina estaria na intersecção de estatística com computação. Ele define a área entre ciência da computação e a área fim como sendo perigosa, pois apesar do pesquisador ter o ferramental para desenvolver sua análise, ele corre o risco de errar na interpretação por não entender as características e limites das ferramentas estatísticas usadas.

Figura 2.1: Diagrama de Venn mostrando as habilidades necessárias para um cientista de dados (CONWAY, 2013).

Com as técnicas da ciência de dados, podemos responder perguntas como “Qual a melhor rota para chegar até meu trabalho?”, “Vai chover hoje?”, “Quais são minhas chances de desenvolver câncer nos próximos 10 anos?”. Neste trabalho, estamos interessados nas técnicas de desenvolvimento de modelos preditivos que, baseados nos dados e conhecimento que temos do assunto, determinam a probabilidade de um resultado.

2.1.1 Modelos preditivos

Geisser (1993) define modelagem preditiva como sendo “o processo pelo qual um modelo é criado ou escolhido para tentar melhor prever a probabilidade de um resultado”. Já Kuhn e Johnson (2013) definem como “o processo de desenvolvimento de uma ferramenta ou modelo matemático que gere uma previsão precisa”.

Modelos preditivos fazem parte do dia a dia da sociedade atual. O Google os utiliza

para interpretar o que seus usuários estão buscando; o Netflix usa para recomendar filmes; corretoras de valores usam para definir que ações comprar ou vender; seguradoras usam para definir qual o risco e, consequentemente, o preço do seguro de um carro, entre outros (LEVY, 2010).

Existem duas principais categorias de problemas que podem ser resolvidos por modelos preditivos: regressão e classificação. A diferença entre eles está no tipo de resposta que queremos prever, se é contínua ou categórica, o que influencia nos tipos de modelos que podem ser usados e na forma de avaliá-los. Por exemplo, ao analisarmos modelos que preveem o preço de um imóvel a partir da sua área, buscamos o que chegue mais próximo do valor real[1]. Já ao avaliar modelos que classifiquem imagens dos dígitos 0 a 9, só nos interessamos na classificação correta; dois modelos que identifiquem a imagem do número 1 como sendo do número 3 e 8 estão igualmente errados. O que classificou a imagem como 3 não está "menos errado" do que a classificou como 8 (KUHN; JOHNSON, 2013; ZUMEL; MOUNT, 2014). O foco deste trabalho é classificação. A seguir, apresentarei os modelos utilizados.

2.1.1.1 Modelo Linear Generalizado (GLM)

O Modelo Linear Generalizado (GLM) é uma técnica para encontrar uma função que, a partir das variáveis independentes, preveja as variáveis dependentes. Ele é uma generalização da regressão linear, onde as variáveis dependentes podem seguir uma distribuição diferente da normal, permitindo assim seu uso com variáveis binárias, por exemplo (KUHN; JOHNSON, 2013).

2.1.1.2 *Support Vector Machine* (SVM)

O *Support Vector Machine* (SVM) é um modelo de aprendizagem supervisionada. Ele funciona mapeando as observações em um espaço n-dimensional, e encontrando o hiperplano que

[1]Esta é uma simplificação. Existem diversas outras características a serem analisadas ao escolher um modelo, como facilidade de interpretação, velocidade de execução, entre outras. Aqui consideramos que a única diferença entre os modelos seja o resultado previsto.

divide as classes e está na distância máxima de todos os pontos. A função que define esse mapeamento se chama "kernel". Dependendo da escolha do kernel, o SVM pode aprender relações lineares ou não-lineares (KUHN; JOHNSON, 2013).

2.1.1.3 Árvore de decisão

Árvores de decisão podem ser pensadas como fluxogramas. Em cada nó da árvore há uma pergunta como, por exemplo: "qual a nacionalidade da pessoa?". Essas perguntas vão sendo respondidas até chegar a uma folha, onde estará o valor previsto (KUHN; JOHNSON, 2013).

2.1.1.4 *Random Forest* (RF)

O *Random Forest* (RF) é um modelo de aprendizagem supervisionada que pode ser usado tanto para regressão, quanto para classificação. Ele funciona retornando a moda (na classificação) ou média (na regressão) dos resultados de um conjunto de árvores de decisão, cada uma treinada em um subconjunto aleatório das variáveis independentes.

Ao usar diversas árvores de decisão, treinadas com subconjuntos diferentes dos preditores, o RF diminui a chance do modelo se tornar *overfit*, aumentando sua performance (KUHN; JOHNSON, 2013).

2.1.1.5 *Stochastic Gradient Boosting* (GBM)

O *Stochastic Gradient Boosting* (GBM) é um modelo que une um conjunto de modelos de baixa performance (normalmente árvores de decisão) a um único de alta performance. Cada submodelo m é treinado a partir dos resíduos do modelo $m-1$. Formalmente, temos:

Seja F o modelo final, composto por M submodelos, que preveja os valores $\hat{y} = F(x)$. A cada etapa $1 \leq m \leq M$, um modelo imperfeito F_m é treinado com uma amostra aleatória sem reposição dos dados. Na etapa seguinte, o método adiciona um novo modelo h de forma que $F_{m+1} = F_m(x) + h(x)$. Caso F_{m+1} fosse perfeito, isso significaria que $F_{m+1} = F_m(x) + h(x) = y$, o que é equivalente a $h(x) = y - F_m(x)$. Assim, o modelo $h(x)$ é treinado buscando prever o residual $y - F_m(x)$. Essa técnica é conhecida como *boosting*

(KUHN; JOHNSON, 2013).

2.1.1.6 C5.0

O C5.0 é um modelo baseado no C4.5 que gera estimativas a partir de árvores ou regras de decisão. Ele se diferencia de outros modelos que usam árvores de decisão pelas técnicas usadas para diminuir a complexidade das árvores, diminuindo a probabilidade de se tornar *overfit*. À semelhança do GBM, ele também usa técnicas de *boosting*.

Regras de decisão se diferenciam de árvores ao permitirem múltiplos critérios em cada etapa. Por exemplo, uma regra poderia ser: "o paciente está com febre *e* dor de cabeça?" (KUHN; JOHNSON, 2013).

2.1.1.7 Rede neural (NNET)

Uma rede neural (NNET) é um modelo inspirado no funcionamento do cérebro dos animais. Ele é dividido em camadas com um conjunto de "neurônios". Os neurônios recebem um valor da camada anterior, fazem algum processamento e passam o novo valor para todos os neurônios da camada posterior. Na primeira camada, as entradas são as variáveis independentes, e na última, as saídas são as previsões do modelo.

Tanto o número de camadas quanto o número de neurônios podem ser modificados, dependendo dos padrões dos dados e objetivos do modelo (KUHN; JOHNSON, 2013).

2.1.2 Avaliando performance em modelos de classificação

Uma das formas mais comuns de descrever a performance de um modelo de classificação é através de uma "matriz de confusão". A Tabela 2.1 mostra um exemplo de matriz de confusão com duas categorias: "positivo" e "negativo". Na diagonal principal estão os valores classificados corretamente, e fora dela estão os erros. Dessa matriz podemos extrair diversas métricas, como:

Previsto	Observado	
	Positivo	Negativo
Positivo	Verdadeiro positivo (VP)	Falso positivo (FP)
Negativo	Falso negativo (FN)	Verdadeiro negativo (VN)

Tabela 2.1: Matriz de confusão para um problema com duas classes: "positivo" e "negativo".

Número de amostras positivas:

$$P = VP + FN \tag{2.1}$$

Número de amostras negativas:

$$N = VN + FP \tag{2.2}$$

Acurácia: Proporção de predições corretas

$$ACC = \frac{VP + VN}{P + N} \tag{2.3}$$

Sensibilidade: Proporção de elementos da classe positiva classificados corretamente em relação ao total de elementos positivos

$$SENS = \frac{VP}{VP + FN} \tag{2.4}$$

Especificidade: Proporção de elementos da classe negativa classificados corretamente em relação ao total de elementos negativos

$$ESPEC = \frac{VN}{FP + VN} \tag{2.5}$$

Dessas métricas, a com interpretação mais simples é a acurácia (equação 2.3), que mede o percentual de predições corretas. Entretanto, ela tem algumas limitações. Primeiramente, ela não leva em consideração o tipo do erro. Em diversos problemas, o custo de um falso positivo é diferente do de um falso negativo. Por exemplo, num sistema de classificação de *spam*[2], é preferível que o usuário veja uma propaganda (falso negativo) a que deixe de receber um e-mail importante (falso positivo). Além disso, ela também não considera a frequência natural de cada classe. Se, por exemplo, 90% dos e-mails que eu receba não sejam *spam*,

[2]E-mails não solicitados.

um modelo que simplesmente classifique todo e-mail como não sendo *spam* terá 90% de acurácia (KUHN; JOHNSON, 2013).

O *Kappa* é uma métrica que leva em consideração as proporções das classes. Ela pode assumir valores entre -1 e 1, sendo que 1 indica que todas as predições foram corretas, -1 indica que todas foram erradas, e 0 que o modelo tem uma performance igual a um aleatório. A equação 2.6 mostra como calculá-la, onde ACC é a acurácia observada e $ACC_{esperada}$ é a acurácia esperada (COHEN, 1960).

$$Kappa = \frac{ACC - ACC_{esperada}}{1 - ACC_{esperada}} \tag{2.6}$$

Em geral, o resultado do uso de modelos de classificação é a probabilidade de pertencer a cada categoria. Por exemplo, um modelo pode, ao analisar um e-mail, determinar que ele tem 75% de chances de ser *spam* e, consequentemente, 25% de chances de não ser. Para transformar essa probabilidade em uma categoria, define-se um ponto de corte (KUHN; JOHNSON, 2013).

Grande parte das ferramentas usa 50% por padrão, que geralmente é uma escolha interessante, caso o custo de um falso positivo e falso negativo seja o mesmo, e as categorias sejam balanceadas. Caso contrário, é recomendável testar outros valores. Uma ferramenta para analisar a sensibilidade e especificidade de um modelo usando diferentes pontos de corte é a curva ROC (ALTMAN; BLAND, 1994; BROWN; DAVIS, 2006; FAWCETT, 2006).

A Figura 2.2 mostra um exemplo de uma curva ROC com os valores discriminados em dois pontos: 0,5 e 0,614. A partir desse gráfico, percebemos a troca entre sensibilidade e especificidade (quanto maior um, menor o outro), e podemos escolher o melhor ponto de corte para o modelo em questão levando em consideração nossos objetivos.

Além de auxiliar na escolha do ponto de corte, a curva ROC também pode ser usada para avaliar e comparar modelos desenhando-os num mesmo gráfico ou comparando a área sob suas curvas (AUC). O modelo na Figura 2.2 possui AUC 0,787. Quanto mais próxima a curva está do canto superior esquerdo, maior é seu AUC e, em geral, maior é sua performance. O modelo perfeito tem AUC 1, já um modelo aleatório tem AUC 0,5 e segue na diagonal do gráfico. O que caracteriza um AUC bom ou ruim varia de caso a caso, mas Hosmer Jr.,

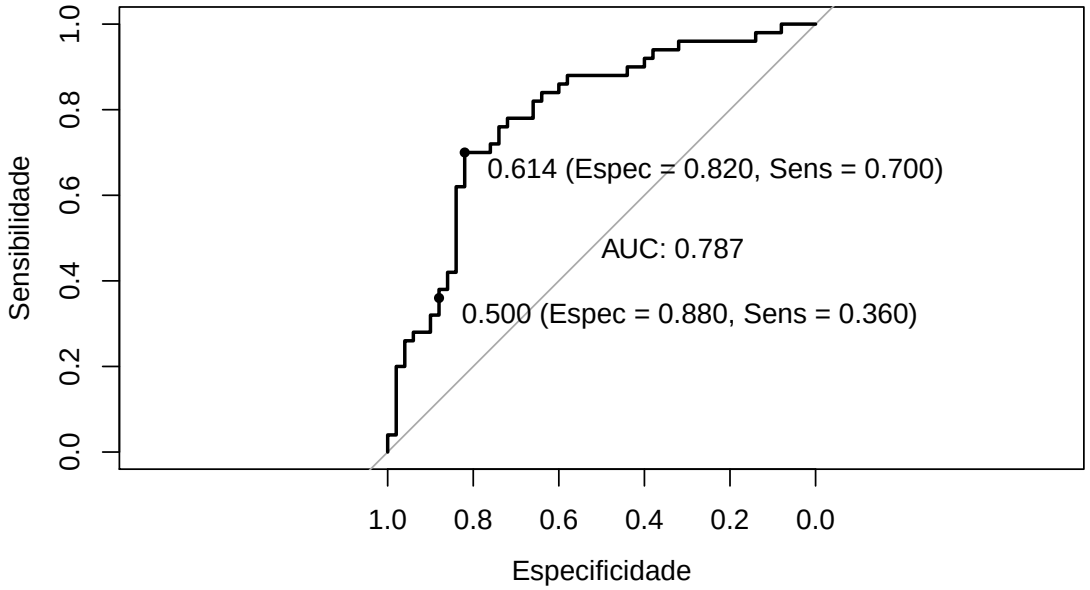


Figura 2.2: Curva ROC de um modelo mostrando a área sob a curva (AUC) e dois pontos de corte: 0,5, o mais comum, e 0,614, o mais próximo do canto superior esquerdo. O gráfico mostra também os valores de especificidade (Espec) e sensibilidade (Sens) nos dois pontos.

Lemeshow e Sturdivant (2013) definiram uma regra geral para avaliação descrita na Tabela 2.2.

2.2 Teoria espacial do voto

É comum usar conceitos espaciais para identificar o posicionamento de atores políticos. Por exemplo, dizemos que um deputado é de esquerda, enquanto outro é de direita. A teoria

AUC	Valor
0,5	Não há discriminação
Entre 0,7 e 0,8	Discriminação aceitável
Entre 0,8 e 0,9	Discriminação excelente
Acima de 0,9	Discriminação excepcional

Tabela 2.2: Regra geral para avaliação da área sob a curva ROC (AUC) de Hosmer Jr., Lemeshow e Sturdivant (2013).

espacial do voto é uma formalização dessas ideias. Ela parte do pressuposto que as preferências individuais e as políticas podem ser interpretadas como pontos em um espaço euclidiano. Um parlamentar então votaria na política mais próxima da sua preferência, chamada de ponto ideal (LEONI, 2002).

As dimensões desse espaço correspondem às áreas temáticas relevantes para os atores, que podem ser muitas. Por exemplo, proteção do meio ambiente, comércio exterior, distribuição de renda, etc.

Consideremos um exemplo onde 5 parlamentares votam em um projeto de Lei que propõe a redução da maioridade penal de 18 para 16 anos. Suponhamos também que cada um tem uma única preferência (do inglês, *single-peakedness*), que é o seu ponto ideal, e vota sinceramente de acordo com ela, e que suas preferências são simétricas. Isto é, dadas duas escolhas a uma mesma distância do ponto ideal de um parlamentar, ele será indiferente a qualquer uma delas.

Graficamente, temos a Figura 2.3, onde as curvas representam as preferências dos legisladores sobre o assunto dessa votação, com a altura representando a intensidade da preferência. Os pontos ideais estão nos picos de cada curva. Há também três linhas, a azul representando o resultado de um voto "sim", a vermelha o de um voto "não" (isto é, o *status quo*), e a preta representa a "linha de corte" dessa votação. As linhas de corte passam pelo ponto médio entre as duas alternativas em votação, separando os parlamentares que deverão votar sim dos que votarão não. Nesse exemplo, as escolhas são entre uma maioridade penal de 16 ou 18, logo a linha de corte passa em $\frac{16+18}{2}$, ou seja, 17. Caso o ponto ideal de um parlamentar esteja sobre essa linha, ele é indiferente às alternativas.

Há diversos modelos criados para estimar esses valores, como o NOMINATE, W-NOMINATE, DW-NOMINATE, que são paramétricos; o *Optimal Classification*, que é não-paramétrico, e modelos baseados em estatística Bayesiana, como o IDEAL. O foco deste trabalho é o W-NOMINATE. (POOLE; ROSENTHAL, 1985; POOLE, 2000; POOLE, 2005; JACKMAN, 2000; CLINTON; JACKMAN; RIVERS, 2004)

O W-NOMINATE faz parte da família de modelos NOMINATE, desenvolvidos por Poole e Rosenthal (1985). Eles são baseados no modelo de utilidade aleatória de McFadden

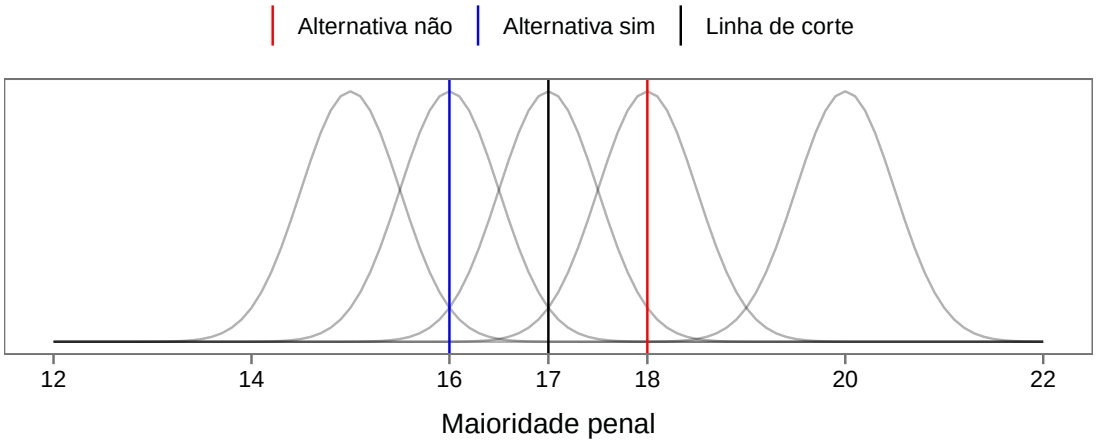


Figura 2.3: Preferências de 5 deputados na votação sobre a redução da maioridade penal de 18 para 16 anos. Parlamentares a esquerda da linha de corte nos 17 anos votam sim, e os a direita votam não. O deputado com preferência sobre a linha de corte é indiferente às duas opções. Neste exemplo, é ele quem desempata a votação.

(1976), que define que a utilidade para um legislador em votar de certa forma é formada por dois componentes: um determinístico e outro estocástico (erro). Sendo s o número de votações indexadas por $k = 1, ..., s$; p o número de legisladores indexados por $i = 1, ..., p$, e; q o número de votações indexados por $j = 1, ..., q$, a utilidade para o legislador i da consequência política do resultado "sim" da votação j é definida como:

$$U_{ijy} = u_{ijy} + \epsilon_{ijy} \tag{2.7}$$

onde u_{ijy} é a parte determinística e ϵ_{ijy} é a parte estocástica da função utilidade U_{ijy}. Se não houver erro, o legislador vota sim se $U_{ijy} > U_{ijn}$, e não se $U_{ijy} < U_{ijn}$, onde y e n representam as consequências de um voto sim ou não, respectivamente. Caso $U_{ijy} = U_{ijn}$, ele é indiferente ao resultado. Com o erro, a diferença entre as utilidades é:

$$U_{ijy} - U_{ijn} = u_{ijy} - u_{ijn} + \epsilon_{ijy} - \epsilon_{ijn} \tag{2.8}$$

logo, um legislador vota sim se:

$$u_{ijy} - u_{ijn} > \epsilon_{ijy} - \epsilon_{ijn} \tag{2.9}$$

ou seja, se a diferença das suas utilidades for maior que a diferença dos erros. No W-NOMINATE, considera-se que a parte determinística u_{ijy} segue a distribuição normal, com os valores concentrados ao redor do ponto ideal do legislador, e rapidamente tendendo a zero ao se afastar para ambos os lados. Com isso, a equação 2.7 se torna:

$$U_{ijy} = \beta \exp^{\left(-\frac{1}{2}\sum_{k=1}^{s} w_k^2 d_{ijky}^2\right)} + \epsilon_{ijn} \tag{2.10}$$

onde β é uma constante que define o peso da parte determinística na equação; w_k é o peso da dimensão k, e; d_{ijky}^2 é a distância ao quadrado na dimensão k do legislador i (X_{ik}) da consequência do resultado sim da votação j (O_{jky}):

$$d_{ijky}^2 = (X_{ik} - O_{jky})^2 \tag{2.11}$$

O método então estima os valores de O_{jky}, O_{jkn}, X_{ik}, β e w_k.

2.2.1 Comparando pontos ideais ao longo do tempo

O principal desafio, ao comparar mudanças nos pontos ideais ao longo do tempo, é distinguir alterações causadas por mudanças na agenda legislativa das causadas por mudanças no posicionamento dos parlamentares (BAILEY, 2007). Em outras palavras, se o ponto ideal de um deputado federal passa de 0,3 para -0,2 de um ano para o outro, como descobrir se isso representa uma mudança real de posicionamento ou é somente reflexo da diferença na agenda legislativa dos dois períodos?

Segundo Shor, Berry e McCarty (2010), todos os esforços para resolver esse problema usam "pontes", que podem ser parlamentares cujo posicionamento assume-se ter se mantido estável durante o período de interesse, ou projetos de Lei que foram votados em mais de um momento (nas duas casas legislativas em um sistema bicameral, por exemplo). O primeiro é

mais usado para comparar as mudanças dos posicionamentos dos parlamentares, enquanto o segundo permite unir pessoas que não votaram juntas em um mesmo mapa espacial[3]. Poole (2005) propõe duas formas para estimar pontos ideais usando pontes.

Na primeira forma, batizada de *pooled scaling* por Shor, Berry e McCarty (2010), dividimos os votos dos parlamentares que queremos mensurar em dois parlamentares "virtuais", um com os votos antes e outro com os votos depois da data de interesse. Unimos esses parlamentares virtuais com os parlamentares-ponte, que possuem um registro único, em uma tabela individual e executamos o algoritmo de estimação dos pontos ideais. Ao final, teremos dois pontos para cada parlamentar de interesse e um ponto para os pontes. Na segunda forma, que Shor, Berry e McCarty (2010) chamam de *linear mapping*, os pontos ideais são estimados separadamente em cada período e conectados com o conjunto de pontos dos parlamentares-ponte através de uma regressão. Ambas formas devem gerar resultados similares, mas a segunda é computacionalmente mais simples, o que pode ser essencial, dependendo da quantidade de parlamentares e votações e votos em estudo.

Poole (2005) ainda descreve uma terceira forma, similar a *pooled scaling* descrita acima, que usa para testar se os senadores norte-americanos mudaram de comportamento nos últimos dois anos de seus mandatos, antes de concorrer à reeleição. Neste caso, ele quer calcular a mudança de comportamento de todos os legisladores em dois momentos, numa mesma legislatura. Se usássemos o *pooled scaling* diretamente, precisaríamos escolher alguns parlamentares como pontes que, por definição, não teriam mudado de comportamento. Ao invés disso, ele segue o seguinte processo:

1. Para cada parlamentar, faça:

    (a) Transforme-o em dois parlamentares "virtuais", um com o conjunto de votos antes, e outro com os depois da data de interesse;

    (b) Calcule os pontos ideais, usando o conjunto de todos parlamentares, com a linha referente ao parlamentar dessa iteração substituída pelos dois parlamentares "virtuais" definidos no passo anterior;

[3]Por exemplo, Shor, Berry e McCarty (2010) colocam todos os legisladores de 11 estados americanos e do congresso federal em um período que varia entre 7 e 15 anos, dependendo do estado, em um mesmo mapa espacial.

(c) Guarde os resultados dos dois parlamentares "virtuais". A diferença entre suas posições representa a mudança de comportamento desse parlamentar.

2. Calcule medidas de incerteza para as estimativas.

Ao final, ele tem dois pontos para cada senador: um representando sua posição nos primeiros 4, e o outro nos últimos 2 anos da legislatura. Note que, como as matrizes de votações usadas para gerar os mapas espaciais são diferentes, eles não são estritamente comparáveis. Apesar disso, Poole (2005) considera que é seguro compará-los, já que eles possuem o mesmo conjunto de votos, com uma única diferença: em cada um dos mapas, um parlamentar distinto é dividido em dois.

2.3 Considerações finais

O objetivo deste capítulo foi fornecer um conjunto mínimo de conhecimento necessário para a compreensão do restante deste trabalho. Para isso, foram descritos o processo de desenvolvimento e avaliação de modelos preditivos, com ênfase em modelos de classificação. Em seguida, foi explicada brevemente a teoria espacial do voto, em especial a técnica W-NOMINATE.

No próximo capítulo, serão apresentados alguns trabalhos relacionados aos temas desta dissertação, analisando suas metodologias e resultados em comparação com o deste trabalho.

Capítulo 3

Trabalhos Relacionados

3.1 Análise do comportamento parlamentar

No final da década de 80 e na de 90, diversos autores viam o sistema político brasileiro como sendo altamente instável, com um Poder Executivo fraco, obrigado a negociar com cada parlamentar individualmente para executar seu plano de governo. Os parlamentares, preocupados somente com seus interesses individuais e regionalistas, seriam indisciplinados, levando o país à ingovernabilidade (ABRANCHES, 1988; LAMOUNIER, 1994; MAINWARING, 2001; AMES, 2003).

Contrários a esse diagnóstico, Limongi e Figueiredo (1995) descreveram um cenário muito diferente. O Executivo, por ter domínio sobre a agenda legislativa, seria capaz de executar seu plano de governo, negociando diretamente com os partidos políticos, que seriam capazes de disciplinar seus parlamentares a votarem conforme suas indicações. Esse diagnóstico foi feito a partir da análise das votações nominais na Câmara dos Deputados no período de 1989 a 1994. Nessa análise, os autores calcularam a disciplina interna de cada partido, usando o índice de Rice (IR)[1]. Todos os partidos apresentaram níveis de disciplina relativamente altos, com o menos disciplinado, o PDS, atingindo um IR 75,70, e o mais

[1]O índice de Rice (IR) varia de 0 a 100, e é calculado subtraindo a proporção dos votos minoritários dos majoritários. Ele é 0 quando o partido está dividido, com metade votando sim e a outra metade votando não, e é 100 quando todos os parlamentares votam da mesma forma (RICE, 1924).

disciplinado, o PT, com IR 95,96.

Essas duas correntes foram tão distintas que Power (2010) as dividiu em "pessimistas" e "otimistas".

Figueiredo e Limongi continuaram analisando dados relacionados ao mesmo tema, focando períodos e características distintos, que confirmaram o prognóstico. A coletânea desses artigos gerou o livro "Executivo e Legislativo na nova ordem constitucional" (FIGUEIREDO; LIMONGI, 2001).

Também usando métricas de disciplina, Cheibub, Figueiredo e Limongi (2009) concluem que o Congresso brasileiro, ao contrário do previsto pelos pessimistas, é altamente centralizado, com partidos e seus líderes capazes de disciplinarem seus parlamentares, evitando que estes ajam somente em busca de benefícios para si mesmo e seus redutos eleitorais.

Além de cientistas políticos, jornalistas também têm usado técnicas semelhantes para análise das votações. O Grupo Estado, dono do jornal Estado de São Paulo (o Estadão), lançou em maio de 2012 a ferramenta Basômetro[2], um site interativo que permite a análise de como os parlamentares, individualmente e agregados por partido, votaram, com dados a partir do primeiro governo de Lula em 2003. Além da análise de cada votação, o Basômetro também calcula a taxa de governismo[3] dos partidos e parlamentares, mostrando suas flutuações em cada mês (ESTADÃO, 2012).

Dantas, Toledo e Teixeira (2014) organizaram uma coletânea de artigos escritos baseados em análises feitas com o Basômetro, gerando o livro "Análise política & jornalismo de dados".

Apesar de serem de interpretação e cálculo mais simples, métricas de disciplina têm alguns problemas. Em primeiro lugar, pela distância entre valores não ser uniforme (por exemplo, a distância entre 40% e 50% não é necessariamente igual a entre 50% e 60%), eles só ordenam os parlamentares. Há também um baixo valor de números possíveis. Em p votações, há $p + 1$ valores possíveis. Por exemplo, em 3 votações, os parlamentares só

[2]Disponível em `http://estadaodados.com/basometro/`.

[3]A taxa de governismo é calculada como o percentual de vezes que os parlamentares votaram de acordo com a posição do governo.

podem assumir os valores 0%, 33%, 66% e 100%. Por fim, o índice é calculado em relação a algum ponto específico. Por exemplo, a taxa de governismo é calculada com base no voto do governo, que normalmente é definido pela indicação ou voto do líder do governo. Assim, ninguém pode ser mais governista do que o próprio líder, o que é um pressuposto que precisa ser testado (POOLE, 2005; MCCARTY, 2011; IZUMI, 2013). Para evitar essas limitações, diversos autores usam modelos espaciais de votação (ver Seção 2.2).

No Brasil, o primeiro autor a usar esses modelos foi Leoni (2002), que analisou o posicionamento dos deputados federais entre 1991 e 1998, usando o W-NOMINATE. Desposato (2005) recorreu ao mesmo método para analisar o efeito da mudança partidária no comportamento dos parlamentares. Zucco (2009) também usa o W-NOMINATE, dessa vez para entender os fatores que influenciam o comportamento dos parlamentares, constatando que a ideologia não explica completamente seus comportamentos. Ele encontrou indicativos de que o presidente da República é um importante influenciador, especialmente através da distribuição de cargos e recursos.

Freitas, Izumi e Medeiros (2012) analisam o significado da primeira dimensão, estimada usando W-NOMINATE, na Câmara dos Deputados e no Senado Federal, concluindo que ela está ligada ao conflito governo e oposição. Izumi (2013) repete essa análise no Senado Federal usando outro modelo, o *Optimal Classification*, chegando à mesma conclusão.

Além de técnicas específicas para gerar mapas espaciais de votação, alguns autores usaram técnicas de redução de dimensionalidade mais gerais. Trento e Leite (2013) analisaram os dados da Câmara dos Deputados, Senado Federal e Câmara Municipal de São Paulo, usando Análise de Componentes Principais (ACP)[4], que é uma técnica estatística de redução de dimensionalidade. Eles mapeiam um voto sim ou não com os valores 1 e 0 e transformam o conjunto de votos dos parlamentares em pontos em um espaço n-dimensional, onde cada dimensão representa uma votação. Como o número de votações é alto, é difícil visualizar esses pontos. Assim, usam o ACP para reduzir esses pontos a um espaço bidimensional, mantendo ao máximo suas posições relativas. Dessa forma, quanto mais próximo estiverem dois parlamentares (ou partidos), mais vezes eles votaram da mesma forma. O resultado

[4]Do inglês, *Principal Component Analysis* (PCA).

dessa pesquisa foi o Radar Parlamentar[5], um site interativo no qual é possível visualizar os diversos gráficos gerados, com as posições dos legisladores e partidos em cada ano.

Usando uma técnica semelhante, a *Multiple Correspondence Analysis* (MCA), Andrade e Monteiro (2015) colocam os deputados federais em um espaço bidimensional, e mostram onde estaria o presidente da Câmara dos Deputados, Eduardo Cunha (PMDB/RJ). Como ele não vota, consideraram seu posicionamento como sendo igual ao do líder do PMDB na Câmara no período em questão. Eles também analisaram a similaridade dos deputados com o Eduardo Cunha através do percentual de votos que deram iguais a ele, batizando essa métrica de "Cunhômetro".

Apesar dessas técnicas estatísticas como a ACP e MCA serem similares a técnicas como o W-NOMINATE, não encontramos trabalhos que validem seu uso em dados de votação, o que pode dificultar a interpretação dos seus resultados.

3.1.1 Análise da mudança de comportamento parlamentar

O problema básico ao comparar mudanças de comportamento ao longo do tempo é distinguir as alterações causadas por mudanças na agenda legislativa, das causadas por uma real mudança das preferências dos parlamentares (BAILEY, 2007). Em outras palavras, se, em um momento, dois parlamentares votaram 90% das vezes da mesma forma e, em outro momento, eles votaram 50% das vezes, como definir se essa mudança se deu porque eles mudaram seus posicionamentos, ou simplesmente porque eles concordavam nas votações do primeiro momento, mas não nas do segundo?

Segundo Shor, Berry e McCarty (2010), todos os esforços para resolver esse problema usam "pontes", que são parlamentares que estiveram presentes em ambos os momentos, e que o analista assume que não mudaram de comportamento nesses momentos. Podemos citar como exemplos parlamentares que foram eleitos para mais de uma legislatura, ou que mudaram de Casa (deputados federais que se tornam senadores). Projetos de Lei também podem ser usados como ponte, caso eles tenham sido votados nas instituições ou períodos de interesse.

[5]Disponível em `http://radarparlamentar.polignu.org`

Nesse artigo, Shor, Berry e McCarty (2010) usam três tipos de pontes para colocar parlamentares que serviram no nível estadual e federal, em ambas as casas[6] e no tempo. Parlamentares que serviram por múltiplas legislaturas tanto a nível estadual quanto federal servem de ponte entre as legislaturas; legisladores que passam da casa baixa para a casa alta (ou vice-versa) em nível estadual conectam as respectivas casas; e parlamentares que passam do nível estadual para atuar no nível federal conectam o estado com o Congresso.

No Brasil, Leoni (2002) analisa as posições dos deputados federais na 49ª e 50ª legislaturas não usando pontes, mas sim a correlação de Pearson entre os pontos ideais. Ele encontrou correlações altas, em torno de 0,80. Apesar disso, ele reconhece que uma limitação de seu trabalho é que, nos períodos estudados, o governo sempre foi de direita, o que poderia causar essa baixa mudança de pontos ideais. Além disso, como ele não usou nenhuma técnica para diferenciar mudanças de comportamento de mudanças na agenda legislativa (como as pontes), não é possível distinguir a razão desse resultado com segurança.

Desposato (2005), ao analisar o impacto das mudanças de partido no comportamento dos parlamentares, usou o posicionamento dos legisladores que não mudaram de partido como pontes. Izumi (2013) faz uma pesquisa semelhante, mas analisando a mudança de comportamento dos senadores que mudaram de um partido dentro da coalizão governamental para um fora (ou vice-versa), para entender o efeito da coalizão no comportamento dos parlamentares.

O Basômetro e o Radar Parlamentar não diferenciam mudanças de comportamento reais das causadas por mudança da agenda legislativa (ESTADÃO, 2012; TRENTO; LEITE, 2013).

Em geral, trabalhos que usam métricas de disciplina interpretam as razões de mudanças de comportamento usando métodos qualitativos, enquanto os que usam pontos ideais usam métodos quantitativos.

[6] O sistema legislativo norte-americano, ao contrário do brasileiro, é bicameral tanto a nível federal quanto estadual (exceto o estado de Nebrasca, que só possui um Senado estadual).

3.2 Aprendizagem de máquina na Ciência Política

Os trabalhos encontrados que usam técnicas de aprendizagem de máquina no âmbito da Ciência Política dividem-se em duas categorias: os que analisam textos (como discursos) buscando entender seu conteúdo ou o posicionamento do autor, e os que analisam projetos de lei tentando prever seus resultados.

Dos que fazem análise textual, Thomas, Pang e Lee (2006) criaram um modelo que classifica os discursos em debates sobre projetos de lei no congresso estadunidense como sendo contrários ou favoráveis a eles. Quinn et al. (2006) criaram um modelo que extrai tópicos do texto de discursos, validando-o nos dados do 105º ao 107º senado dos Estados Unidos. Yu, Kaufmann e Diermeier (2008) determinam a filiação partidária dos parlamentares estadunidense a partir dos seus discursos. Somasundaran e Wiebe (2010) classificam autores de textos publicados na Internet, identificando seu posicionamento político a partir de suas argumentações. Já Conover et al. (2011) fazem o mesmo com os usuários do Twitter[7], mas a partir da frequência do uso de certas palavras.

Já na previsão do sucesso de projetos de lei, Gerrish e Blei (2011) e Goldblatt e O'Neil (2012) desenvolveram modelos que acertaram mais de 90% dos resultados nos períodos estudados. Yano, Smith e Wilkerson (2012) fizeram uma análise semelhante, mas focando prever os projetos de lei que conseguirão passar das discussões ocorridas em comissões e ser postos em votação. Wang, Varshney e Mojsilovi (2012) preveem os resultados através de uma caminhada aleatória em 3 grafos heterogêneos interconectados, sendo dois representando os legisladores, e um representando projetos de lei. Os legisladores estão interligados caso tenham escrito ao menos um projeto juntos[8], e os projetos de lei interligados pela análise da similaridade textual. Quando um legislador vota em um projeto de lei, um vértice é criado entre seu nó e o do projeto. Ele criou dois grafos para os legisladores como forma de criar dois tipos de ligação entre os legisladores e os projetos de lei: uma para representar votos sim, e outra para votos não.

A empresa americana Fiscal Note (2015) desenvolve ferramentas para monitoramento

[7]Disponível em `http://twitter.com`.

[8]Do inglês, *co-sponsorship*.

legislativo usando técnicas de ciência de dados. Uma delas, a *Prophecy*[9] (do inglês, profecia), supostamente prevê o resultado de projetos de lei com 94% de acurácia. No Brasil, o ELLO, empresa ligada ao CEBRAP, possui um produto similar.

3.3 Considerações finais

Como visto na Seção 3.1, o foco da literatura encontrada é na análise e interpretação do comportamento dos parlamentares, seja usando métricas de disciplina ou modelos espaciais de votação. Na Seção 3.2, foram apresentados trabalhos que aplicam técnicas de aprendizagem de máquina em temas da ciência política, buscando descobrir o posicionamento de pessoas através dos seus textos ou prever o resultado de votações.

Nesse levantamento bibliográfico, não foi encontrado nenhum autor que desenvolva o objetivo deste trabalho: a detecção de mudanças de posicionamento dos legisladores. O que parece ser mais próximo é a pesquisa feita pela empresa americana Fiscal Note (2015) para o desenvolvimento do produto *Prophecy*, mas não foi possível analisar as semelhanças mais a fundo, pois o processo usado por eles não é divulgado.

Nesse sentido, considero a definição da metodologia para o desenvolvimento de um modelo de detecção de mudança de posicionamento dos deputados federais brasileiros como sendo o principal diferencial e contribuição deste trabalho.

[9] Disponível em `https://www.fiscalnote.com/prophecy`.

Capítulo 4

Desenvolvimento do modelo preditivo

Neste capítulo, apresentaremos o processo de criação do modelo preditivo, partindo da coleta e preparação dos dados, onde serão estimadas as mudanças de comportamento, passando pela análise dos dados com objetivo de definir os preditores (variáveis independentes). Ao final, treinaremos 6 tipos de modelos diferentes, escolhendo o final de acordo com a área sob a curva ROC (AUC).

4.1 Coleta dos dados

Os votos e votações foram extraídos diretamente do site da Câmara dos Deputados através de sua *Application Programming Interface* (API)[1] (Câmara dos Deputados, 2015). Para desenvolver o modelo preditivo, precisamos ter acesso a todos os dados na máquina local. Entretanto, até a data de publicação deste trabalho, a API da Câmara só permitia consultas individuais, votação por votação, por isso desenvolvi um programa[2] na linguagem Python que, com o auxílio da biblioteca Scrapy, baixa todas as votações (Python Software Foundation, 2014; HOFFMAN et al., 2014). A Tabela 4.1 mostra o número de deputados federais, votos e votações por legislatura. Em média, temos 624 parlamentares e 332 votos em 478 votações a cada legislatura.

[1] Em tradução livre, "Interface de Programação de Aplicativos".

[2] Disponível em: `https://github.com/vitorbaptista/dados-abertos-camara.gov.br`

Legislatura	Deputados Federais	Votações	Votos
50	631	468	178603
51	624	419	155737
52	614	451	134461
53	606	619	192879
54	644	432	131552

Tabela 4.1: Número de deputados federais e votações por legislatura

A lista das coalizões foi obtida através do banco de dados legislativos do CEBRAP. Ela contém a data de início e fim e a composição partidária de cada coalizão. Como trabalhamos com os parlamentares individualmente, gerei outra lista que contém todas as entradas e saídas dos deputados federais na coalizão dentro de uma mesma legislatura. Por exemplo, apesar do deputado Arlindo Chinaglia (PT/SP) passar de oposição a governo em 2003 com a eleição de Lula, não consideramos isso uma mudança pois ela ocorreu entre legislaturas. Já a saída de Romário (PSB/RJ) da coalizão quando seu partido se tornou oposição em 2013 é considerada, pois ocorreu dentro de uma mesma legislatura.

As coalizões são compostas pelos partidos que controlam ao menos um ministério. Uma nova coalizão é formada em duas situações: i) na mudança de governo, e; ii) na mudança do conjunto de partidos que controlam ministérios. Em outras palavras, uma nova coalizão é formada quando um partido que não possuía nenhum ministério passa a ter ao menos um, quando um partido que controlava algum ministério passa a não ter nenhum, ou quando se inicia um novo governo. Esses critérios foram definidos por Figueiredo (2007), inspirada no trabalho de Müller e Strom (2000).

A Figura 4.1 mostra as 890 entradas e saídas da coalizão que ocorreram durante o período de estudo, agrupadas mensalmente. Há meses que concentram as mudanças de posicionamento, que serão melhor analisadas na Seção 4.3.2. A linha tracejada marca a data da Resolução nº 22.610 do TSE, que em 24/10/2007 alterou o entendimento das regras de fidelidade partidária, tornando-as mais restritas (Tribunal Superior Eleitoral, 2007). Aparentemente, essa mudança alterou os aspectos temporais das mudanças de posicionamento, o que influencia na construção do modelo, mais especificamente na divisão dos dados, tratada

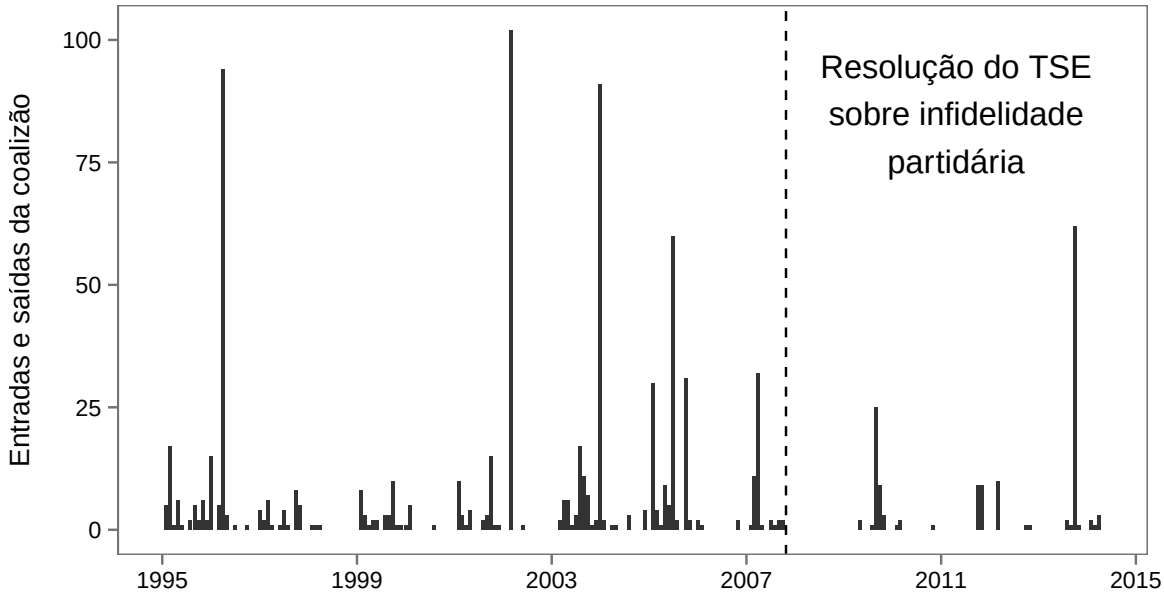


Figura 4.1: Número de deputados federais que mudaram de posicionamento, entrando ou saindo da coalizão entre a 50ª e 54ª legislaturas, agrupados mês a mês.

na Seção 4.4.3.

4.2 Preparação dos dados

O período de análise deste trabalho é composto pela 50ª, 51ª, 52ª, 53ª e 54ª legislaturas, compreendendo os 20 anos de 1995 até o início de 2015, entre o primeiro mandato de Fernando Henrique Cardoso até o início do segundo mandato de Dilma Rousseff. As 48ª e 49ª legislaturas, iniciadas respectivamente em 1987 e 1991, foram excluídas pois, segundo Freitas (2008), elas ocorreram em um período em que parlamentares ainda estavam se acomodando às novas regras advindas da transição à democracia.

A análise se restringe à Câmara dos Deputados por duas razões: facilidade na obtenção dos dados e maior número de parlamentares (513 contra 81 no Senado Federal) (VICENTE, 2012).

A unidade de análise usada é o parlamentar. Apesar da literatura apontar o papel fundamental dos partidos no comportamento dos parlamentares (FIGUEIREDO; LIMONGI,

2001; SANTOS, 2003; FREITAS, 2008), trabalhar com os dados desagregados nos permite perceber movimentações individuais. Por exemplo, quando um parlamentar deixa um partido governista para entrar em um de oposição (ou vice-versa). Considerei todos os parlamentares que votaram ao menos uma vez na Câmara, inclusive os suplentes, por isso o número é maior do que os 513 eleitos por legislatura.

Os deputados podem votar sim, não, nulo, branco, se absterem ou obstruirem (CARNEIRO; SANTOS; NETO, 2013), mas o método W-NOMINATE, usado neste trabalho, só considera votos sim ou não. Isso nos dá duas opções: mapear os outros tipos de votos como sendo a favor ou contra (por exemplo, considerando abstenções como votos contrários), ou ignorá-los. Para evitar possíveis problemas metodológicos ao criar critérios desse tipo, desconsidero votos diferentes de sim ou não.

4.2.1 Estimando a mudança de comportamento

Para estimar a mudança de comportamento, seguiremos a metodologia de Poole (2005) usando o modelo W-NOMINATE (POOLE; ROSENTHAL, 1985) (ver Secão 2.2.1). Esse modelo foi escolhido porque, apesar de diversos outros modelos terem sido desenvolvidos nos 30 anos desde sua criação, ele continua sendo um dos modelos mais usados (POOLE et al., 2011), inclusive no Brasil (LEONI, 2002; DESPOSATO, 2005; ZUCCO, 2009; FREITAS; IZUMI; MEDEIROS, 2012; BERNABEL, 2015).

Ao estimar os pontos ideais, usando o W-NOMINATE, precisamos definir 5 parâmetros: número de dimensões; polaridade(s); critérios de inclusão de parlamentares e votações; número de repetições para estimar o erro; período de análise. Eles serão definidos nas próximas subseções.

4.2.1.1 Número de dimensões

Leoni (2002) analisou o número de dimensões necessárias para explicar o comportamento dos parlamentares na Câmara dos Deputados de 1991 a 1998. Freitas, Izumi e Medeiros (2012) expandiram esse trabalho analisando não só a Câmara, mas também o Senado Federal,

usando os dados do período de 1988 a 2010. Izumi (2013) analisou o Senado de 1989 a 2010 usando outro modelo espacial de votação, o *Optimal Classification*. Exceto durante a 49ª legislatura, onde tivemos dois presidentes (Collor e Itamar Franco), todos concluíram que o congresso é predominantemente unidimensional. Como essa legislatura não será analisada neste trabalho, usarei uma única dimensão.

4.2.1.2 Polaridade

A polaridade define qual é a direção dos eixos das dimensões. No caso de uma única dimensão, o parlamentar escolhido como polaridade será colocado nos valores positivos. Idealmente, deve-se escolher um parlamentar que não tenha mudado de posicionamento no período e que tenha participado do maior número de votações. Como o PT é, historicamente, um dos partidos mais disciplinados do Brasil e nunca mudou de posicionamento durante uma legislatura, sendo sempre oposição até se tornar governo com a eleição de Lula em 2003, escolhi parlamentares filiados a ele como polaridade.

O deputado José Genoíno foi usado nas 50ª, 51ª e 53ª legislaturas, Arlindo Chinaglia na 52ª e Luiz Couto na 54ª. Eles foram escolhidos porque nunca trocaram de partido, ocuparam cargos de liderança e participaram de um grande número de votações nas legislaturas em que foram escolhidos.

4.2.1.3 Critérios de inclusão de parlamentares e votações

Incluímos parlamentares que participaram de, no mínimo, 20 votações, cujo lado minoritário foi responsável por 2,5% ou mais dos votos. Esses são os critérios padrão do W-NOMINATE, seguidos nos trabalhos de Leoni (2002), Zucco (2009), Freitas, Izumi e Medeiros (2012).

4.2.1.4 Períodos de análise

Na análise de uma mudança de comportamento, é preciso definir ao menos dois períodos: um anterior e outro posterior à data de interesse. Por exemplo, para analisar o efeito da saída

do PSB da coalizão em outubro de 2013 no comportamento dos parlamentares, poderíamos definir o período inicial como sendo do início da legislatura até à data da saída do PSB, e o período final como sendo da saída do PSB até ao final da legislatura. Como o objetivo deste trabalho não é analisar um período específico, mas criar um modelo que detecte mudanças na coalizão, não basta definir um, mas sim um conjunto de períodos.

Ao definir esses períodos, é preciso levar em consideração diversos fatores. Períodos muito curtos, como uma semana antes e uma semana depois, podem sofrer com a falta de votações suficientes ou serem demasiadamente influenciados por votações anômalas. Já períodos muito longos, como 5 anos antes e 5 anos depois, dificultam a interpretação dos resultados, pois podem conter diversas mudanças de comportamento. No mínimo, precisamos de um período que contenha votações suficientes de acordo com nossos critérios de inclusão definidos na Seção 4.2.1.3.

Baseado nisso, defini períodos de 12 meses dentro de uma mesma legislatura, divididos em duas partes do mesmo tamanho. Eles iniciam no primeiro dia de cada mês, partindo de 01/fev do primeiro ano da legislatura até ao do último. Por exemplo, na 54ª legislatura, o primeiro período vai de 01/fev/2011 até 01/fev/2012 (dividido em 01/ago/2011), e o último vai de 01/fev/2014 até 01/fev/2015. Existem 37 desses períodos por legislatura, 185 nas 5 legislaturas estudadas neste trabalho. A arbitrariedade dessa escolha é uma limitação deste trabalho, que será melhor analisada na Seção 5.2.

4.2.1.5 Estimativas de erro

Poole (2005) sugere, sem justificar esse número, repetir as estimativas 101 vezes para calcular seu erro. Entretanto, o volume de dados que estamos trabalhando inviabiliza esse número de repetições com os recursos computacionais disponíveis para esta pesquisa[3]. Por isto,

[3]Por exemplo, usando um núcleo de um processador AMD Opteron™ 6238 com 2,6 GHz, o processamento dos pontos ideais de um parlamentar em um período demora cerca de 1h45m. Considerando somente os 513 deputados eleitos por legislatura, o cálculo de todos os 37 períodos levaria aproximadamente 18 meses. Todas as 5 legislaturas analisadas neste trabalho levariam mais de 7 anos. O cálculo é facilmente paralelizável, diminuindo o tempo necessário de acordo com o número de processadores, mas ainda assim não foi possível usar 101 repetições neste trabalho.

escolhi fazer 10 repetições, o que diminuiu o tempo necessário em cerca de 90%.

4.2.2 "Raio de influência" de uma mudança de posicionamento

Até este momento, temos uma tabela onde cada linha representa um parlamentar em um período, e nas colunas temos seus pontos ideais e as respectivas estimativas de erro. Em outras palavras, temos nossas variáveis independentes, agora precisamos determinar a variável dependente: se o parlamentar mudou de posicionamento, entrando ou saindo da coalizão, ou se manteve como estava.

Para isso, precisamos definir qual o período antes de uma mudança de posicionamento no qual o parlamentar começa a mudar de comportamento. Por exemplo, considere que detectamos que um parlamentar, que está fora da coalizão, se aproximou do governo entre o primeiro e o segundo semestres de 2011, mas só entrou na coalizão em 2014. Devemos considerar que essa aproximação em 2011 foi indício da sua entrada em 2014? E se ela tivesse ocorrido em 2012? Em outras palavras, precisamos definir qual o "raio de influência" de uma entrada ou saída da coalizão.

Não há uma resposta única. Alguns parlamentares podem mudar de comportamento anos antes de mudarem de posicionamento, enquanto outros podem não mudar em momento algum. A hipótese fundamental deste trabalho é que eles mudam de comportamento antes de mudar de posicionamento, caso contrário seria impossível prever um a partir do outro. Formalmente, sendo $Data_{coalizão}$ a data de entrada ou saída da coalizão, $Data_{inicial}$ e $Data_{final}$ as datas iniciais e finais do período analisado, $Data_{média}$ a data que divide o período analisado em dois, "antes" e "depois", e $P = Data_{média} - Data_{inicial} = Data_{final} - Data_{média}$ a distância da data inicial até à data média, que é igual à da data média até a data final, então o período em que consideramos que uma mudança de comportamento esteja relacionada à mudança de posicionamento compreende $\left[Data_{média} - \frac{P}{2}, Data_{média} + \frac{P}{2}\right]$. Precisamos definir P.

Para entender melhor, vejamos a Figura 4.2. Nela são mostrados cinco períodos de análise de um parlamentar que mudou de comportamento em $Data_{coalizão} - P$, representado pela mudança da região azul para a verde, com o período de análise em cada momento re-

presentado pela região escurecida entre $Data_{inicial}$ e $Data_{final}$. No primeiro período, 4.2a, a $Data_{final}$ é igual ao momento de mudança de comportamento. Aqui, não consideramos que houve uma mudança de posicionamento, pois o comportamento dentro desse período é constante (região azul). Na Figura 4.2b, a mudança de comportamento ocorre no ponto médio entre a $Data_{média}$ e a $Data_{final}$. A partir daqui começamos a considerar que o parlamentar mudou de posicionamento. Na Figura 4.2c, a $Data_{final} = Data_{coalizão}$, e a $Data_{média}$ se iguala a data da mudança de comportamento em $Data_{coalizão} - P$. Se a escolha de P foi correta para este parlamentar, esse deve ser o período com a mudança de comportamento mais intensa. A Figura 4.2d mostra o último período em que ainda consideramos que houve uma mudança de comportamento causada por uma mudança de posicionamento, pois a partir dele o comportamento em $[Data_{inicial}, Data_{média}]$ começa a se igualar ao em $[Data_{média}, Data_{final}]$. Na última figura, a 4.2e, o comportamento nos dois momentos do período de estudo se igualam.

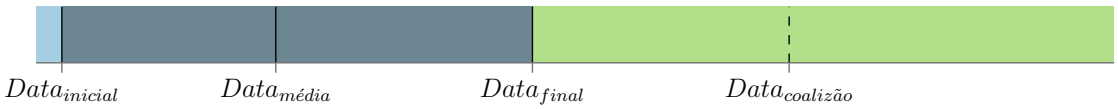
Neste trabalho, defini P como sendo 6 meses. Em outras palavras, os parlamentares começariam a mudar de comportamento a partir de 6 meses antes da definitiva entrada ou saída da coalizão. Por ser uma escolha arbitrária, esta é uma limitação do trabalho. Na Seção 5.2 discorro sobre os possíveis problemas dessa escolha e sugestões sobre como remediá-los em um trabalho futuro.

4.3 Análise dos dados

Segundo Kuhn e Johnson (2013), uma das etapas mais importantes na criação de modelos preditivos é analisar os dados, buscando padrões que poderão ser utilizados para o aumento de sua performance. Nas seguintes subseções, analisaremos as características das estimativas de pontos ideais e seus padrões temporais.

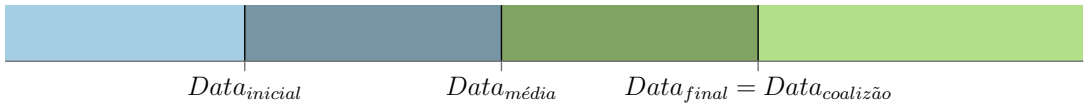
4.3.1 Pontos ideais

O conjunto de dados gerado possui 113.952 estimativas de mudança de comportamento em 185 períodos de 12 meses, divididos em dois subperíodos de mesma duração. Não foi possí-



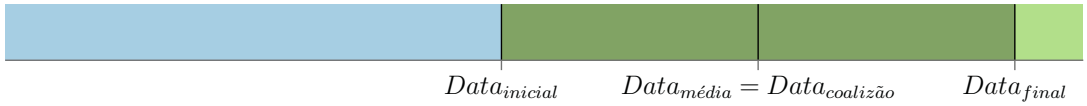
(a) A data final do período de análise é igual ao dia em que o parlamentar mudou de comportamento por causa da futura mudança de posicionamento. Como, dentro desse período atual, supomos que o parlamentar manteve o mesmo comportamento, ainda não consideramos que houve uma mudança.

(b) O raio de influência da coalizão chega à metade do segundo período. A partir daqui começamos a considerar que houve uma mudança de posicionamento, pois o parlamentar esteve com outro comportamento em mais da metade da segunda parte do período.



(c) A $Data_{media}$ se iguala à data de mudança. Este é o período onde esperamos encontrar a maior mudança de comportamento.

(d) A partir daqui, o comportamento começa a se igualar nos dois momentos de análise. Este é o último período em que consideramos que houve uma mudança de posicionamento.



(e) O comportamento nos dois períodos se iguala. Esperamos que a diferença entre os pontos ideais dos dois momentos volte aos níveis da segunda figura.

Figura 4.2: Diversos períodos de análise de um parlamentar que mudou de comportamento em $Data_{coalizao} - P$, representado pela mudança da região branca para a verde.

	50	51	52	53	54
S	0.04	0.04	0.07	0.02	0.03
N	0.96	0.96	0.93	0.98	0.97

Tabela 4.2: Percentual de deputados que mudaram de posicionamento por legislatura.

vel estimar o ponto ideal em um ou ambos subperíodos em 31.04% (35.376) períodos, pois ou parlamentar não participou de votações o suficiente, ou as de que ele participou não passaram nos critérios de inclusão (ver Seção 4.2.1.3). Esses casos foram excluídos, restando ao final 78.576 observações.

Desses valores, 3.98% (3.130) são de períodos em que houve uma mudança de posicionamento, enquanto 96.02% (75.446) não. Por legislatura, existem em média 4% mudanças, com o máximo ocorrendo na 52ª (7.07%) e o mínimo na 53ª (1.55%). A Tabela 4.2 mostra esses dados em todas as legislaturas estudadas.

A Figura 4.3 mostra a densidade dos valores dos pontos ideais em cada legislatura, com o PT estando à direita do gráfico. Na 50ª e 51ª legislaturas, durante o governo do PSDB, a maior parte dos deputados federais encontra-se ao lado esquerdo. Esse quadro muda a partir da eleição de Lula em 2003, na 52ª legislatura, com a maioria dos parlamentares estando mais à direita do gráfico. Na 53ª legislatura, as posições se tornam mais polarizadas, com dois picos claros, um na direita do gráfico e outro, menor, na esquerda. Por fim, durante o primeiro governo de Dilma Rousseff, na 54ª legislatura, os pontos ideais se concentram no centro.

4.3.2 Aspectos temporais

Uma das principais características da migração partidária no Brasil é sua distribuição no tempo (ARAÚJO, 2000; MELO, 2004; FREITAS, 2008). A partir de 1995, ela se concentra nos meses de fevereiro do primeiro e terceiro ano da legislatura e no período pré-eleitoral, que se encerra em outubro do ano anterior às eleições (FREITAS, 2008; BRASIL, 1997). Nesta seção, analisaremos se as mudanças de posicionamento também seguem um padrão temporal.

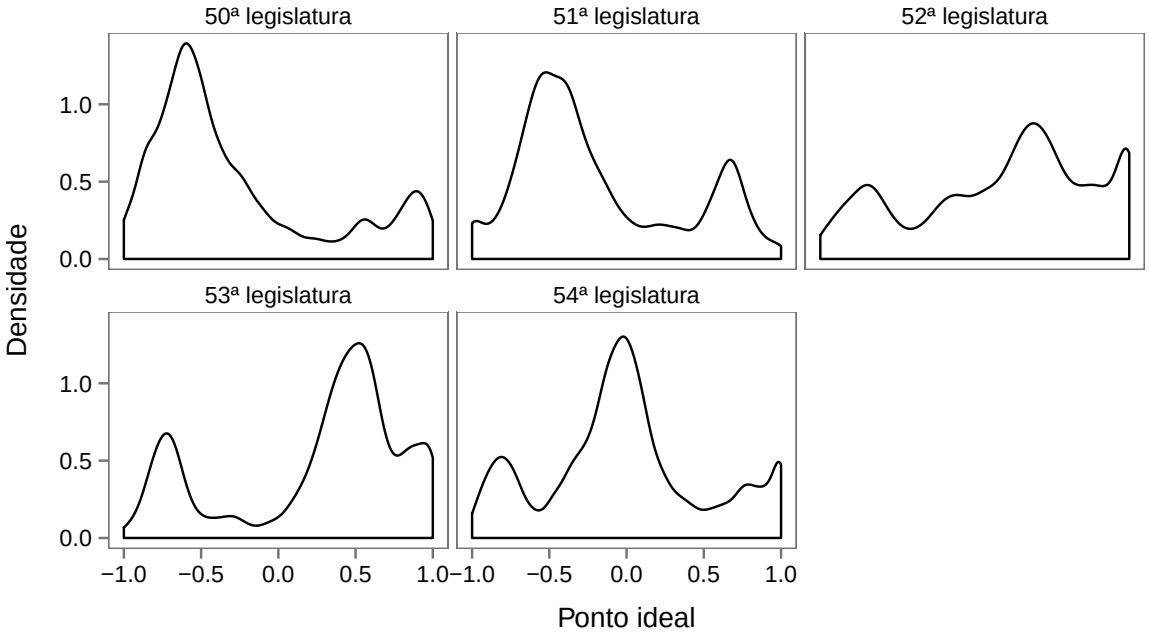


Figura 4.3: Distribuição dos pontos ideais nos 37 períodos de análise a cada legislatura.

Há duas formas de um parlamentar entrar ou sair da coalizão: quando seu partido o faz, ou mudando de partido. Das 890 mudanças de posicionamento no período de estudo, 50.79% (452) ocorreram por migrações e 49.21% (438) pelo partido ter mudado de posicionamento. Consideramos a data de mudança de posicionamento de formas diferentes em cada caso. Quando o partido como um todo entra ou sai da coalizão, consideramos a data de mudança como a data de início da coalizão. Já em uma mudança por migração, usamos a data da primeira votação do parlamentar no novo partido[4].

A Figura 4.4 mostra o número de deputados federais que mudaram de posicionamento entre a 50ª e 54ª legislaturas agregados por mês. Quatro meses, janeiro, março, abril e outubro, concentram 65.06% (579) das mudanças, o que pode ser um indício de que as mudanças de posicionamento sigam um padrão temporal.

Na Figura 4.5, separamos os deputados que mudaram de posicionamento migrando de

[4]Idealmente, a data de mudança por migração deveria ser a data de filiação do parlamentar ao novo partido. Infelizmente, essa informação não está disponível através da API da Câmara para todas as legislaturas analisadas, então usei a data da primeira votação no novo partido. Como, no período de análise, 99% dos parlamentares faltaram no máximo 35 votações seguidas, as duas datas devem ser próximas na maioria dos casos.

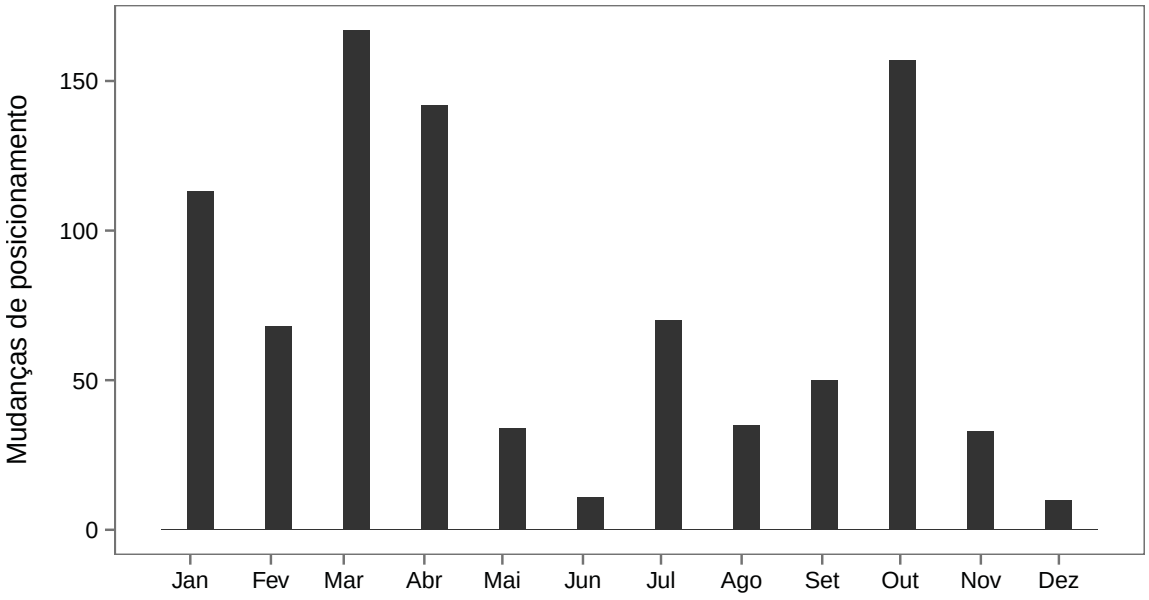


Figura 4.4: Número de deputados que mudaram de posicionamento entre a 50ª e 54ª legislaturas agregados por mês.

partido (migrantes) dos cujo próprio partido mudou de posicionamento (não-migrantes). O padrão temporal dos dois grupos é visivelmente distinto. Os que migraram de partido mudaram de comportamento em média 38 vezes por mês, com 29.87% (135) ocorrendo em outubro. Houve respectivamente 13.27% (60) e 11.28% (51) em março e fevereiro, os segundo e terceiro meses com mais mudanças. Já os parlamentares cujos partidos entraram ou saíram da coalizão mudaram de posicionamento em média 36 vezes por mês, com janeiro, março e maio concentrando 46.8% (205) das mudanças.

Separamos os dados por legislatura na Figura 4.6. Os migrantes continuam com um padrão parecido com, em média, 32.06% das mudanças ocorrendo em outubro, 14.18% em março e 10.48% em fevereiro. Os não-migrantes não seguem um padrão único, mudando a cada legislatura.

Na Figura 4.7, agregamos por ano da legislatura ao invés de mês. A grande maioria das mudanças dos migrantes, 84.07% (380), está concentrada no primeiro e terceiro anos, seguindo o padrão temporal de migração descrito por Freitas (2008). Já os não-migrantes continuam sem um padrão claro. Como só consideramos mudanças dentro da mesma legislatura, só uma ocorreu no primeiro ano, durante o segundo mandato de Lula, quando o

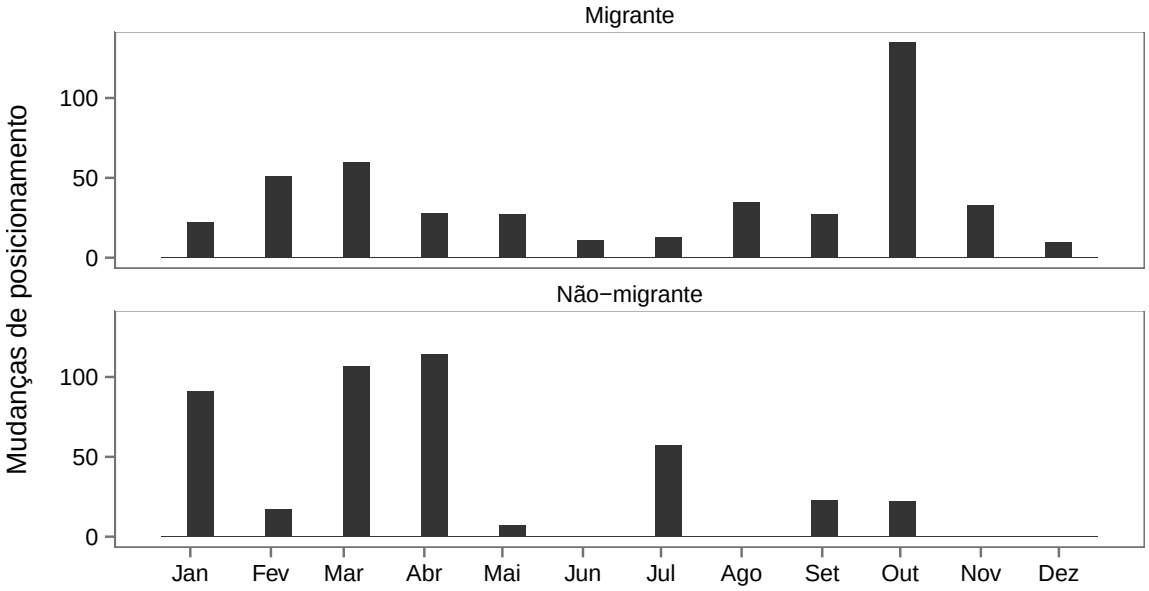


Figura 4.5: Número de deputados migrantes e não-migrantes que mudaram de posicionamento entre a 50ª e 54ª legislaturas agregados por mês.

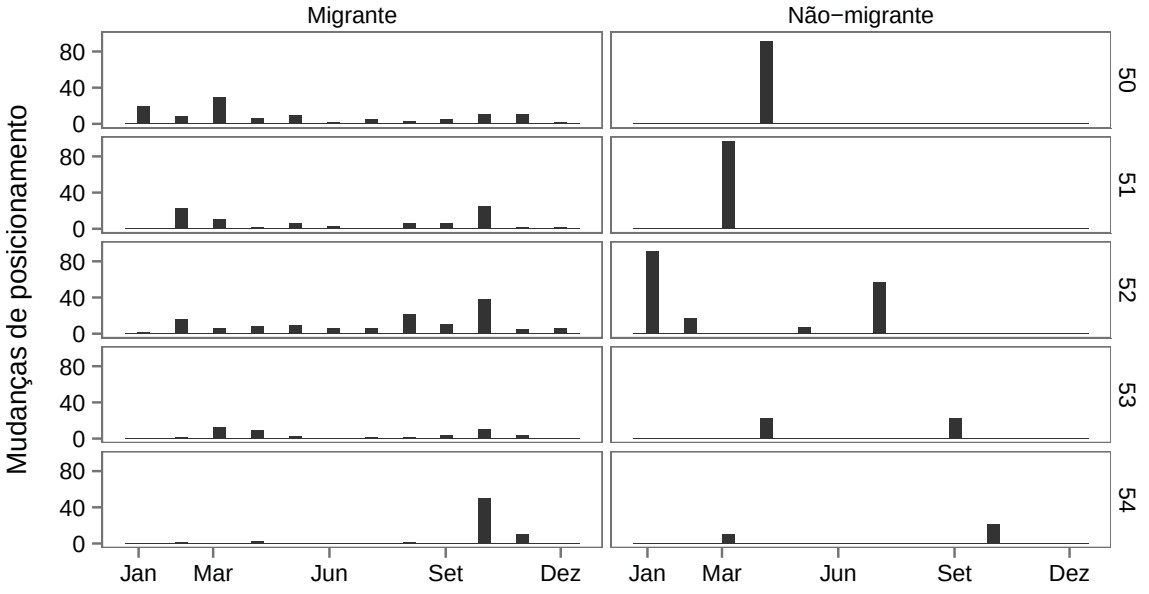


Figura 4.6: Número de deputados migrantes e não-migrantes que mudaram de posicionamento entre a 50ª e 54ª legislaturas agregados por mês e legislatura.

PDT entrou na coalizão em abril de 2007. As outras mudanças estão razoavelmente bem espalhadas do segundo ao último ano das legislaturas, com a maioria ocorrendo no segundo ano.

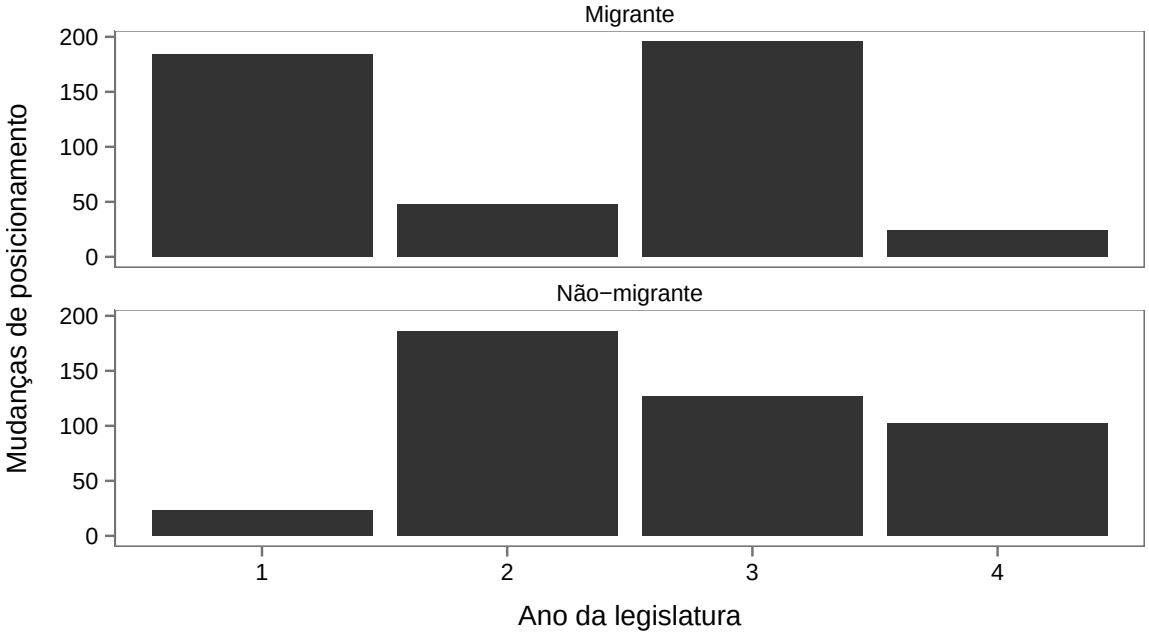


Figura 4.7: Número de deputados migrantes e não-migrantes que mudaram de posicionamento entre a 50ª e 54ª legislaturas, agregados por ano da legislatura.

Separando os dados por mês e ano da legislatura, temos a Figura 4.8. Nos migrantes, 52.55% (103) das mudanças ocorrem em outubro do terceiro ano, o limite para se filiar a um partido para concorrer nas eleições seguintes (BRASIL, 1997). Os não-migrantes continuam sem um padrão claro.

Em suma, os deputados que entram ou saem da coalizão migrando de partido seguem um padrão temporal similar ao das migrações partidárias em geral descrito por Freitas (2008), enquanto os que mudam de posicionamento sem migrar de partido não têm um padrão claro.

4.4 Modelagem

Esta seção trata do desenvolvimento do modelo preditivo. Para isto, inicialmente definiremos as variáveis dependente e independentes e as divisões do conjunto de dados e, em seguida, treinaremos modelos de diversos tipos, buscando o que possui a maior área sob a curva

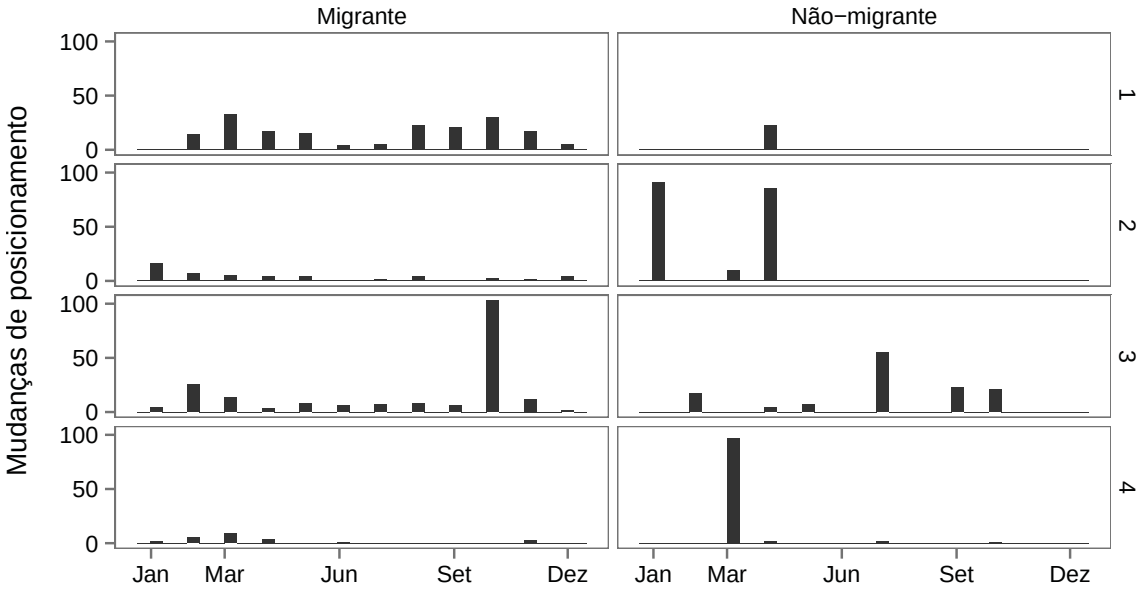


Figura 4.8: Número de deputados que mudaram de posicionamento entre a 50ª e 54ª legislaturas, agregados por mês e ano da legislatura.

ROC (AUC). Encontrado esse modelo, definiremos o ponto de corte para classificação das categorias e estimaremos sua performance final.

4.4.1 Variável dependente

Como variável dependente, que é o que queremos prever, usaremos a mudança ou não de comportamento do parlamentar no período. Ela foi definida a partir dos períodos e "raio de influência" apresentados na Seção 4.2.2 e toma o valor "S", quando houve uma mudança de posicionamento, e "N", quando não houve. Como já visto anteriormente, das 78.576 observações, 3.98% (3.130) são relativas a mudanças de posicionamento, e 96.02% (75.446) não.

4.4.2 Variáveis independentes

A partir dos resultados das análises descritas na Seção 4.3, defini 6 variáveis independentes, sendo 4 relativas às estimativas de pontos ideais e seus erros, e duas relacionadas aos padrões

Variável	Descrição	Natureza
antes	Estimativa do ponto ideal no momento inicial	Contínua
antes.sd	Estimativa de erro do ponto ideal no momento inicial	Contínua
depois	Estimativa do ponto ideal no momento final	Contínua
depois.sd	Estimativa de erro do ponto ideal no momento final	Contínua
mês	Mês na data média do período	Ordinal
ano da legislatura	Ano da legislatura na data média do período	Ordinal

Tabela 4.3: Variáveis independentes usadas na criação dos modelos preditivos.

temporais descritos na Seção 4.3.2. A Tabela 4.3 contém a descrição e natureza de cada uma delas.

A Figura 4.9 mostra a distribuição dos valores das variáveis independentes agrupados pela mudança ou não de posicionamento no período. Visualmente, os padrões são parecidos, mas ao treinar os modelos usei não só essas variáveis, como também as interações entre elas. Dessa forma, eles podem aprender características mais complexas, de difícil visualização.

As variáveis “antes” e “depois” possuem uma correlação alta, de 0,91. Os erros possuem correlação baixa, sendo 0,49 a correlação entre eles, 0,21 a do “antes” com seu erro “antes.sd”, e 0,23 a do “depois” com seu erro “depois.sd”. Ao separar os parlamentares que mudaram de posicionamento, sua correlação entre “antes” e “depois” cai para 0,67, enquanto os outros continuam com correlação 0,91. Essa diferença faz sentido, já que esperamos que quem mudou de posicionamento tenha alterado seus pontos ideais, enquanto os de quem não mudou tenham-se mantido constantes.

4.4.3 Divisão da base de dados

Existem tipos de modelos capazes de aprender a estrutura de um conjunto de dados tão bem que, quando validados com o mesmo conjunto de dados usado para treiná-los, conseguem prever corretamente 100% dos casos. O problema é que eles aprenderam não só os padrões gerais dos dados, mas também seu ruído. Esses modelos estão superajustados (do inglês, *overfit*), e têm normalmente uma baixa performance ao serem usados em novos dados. Para

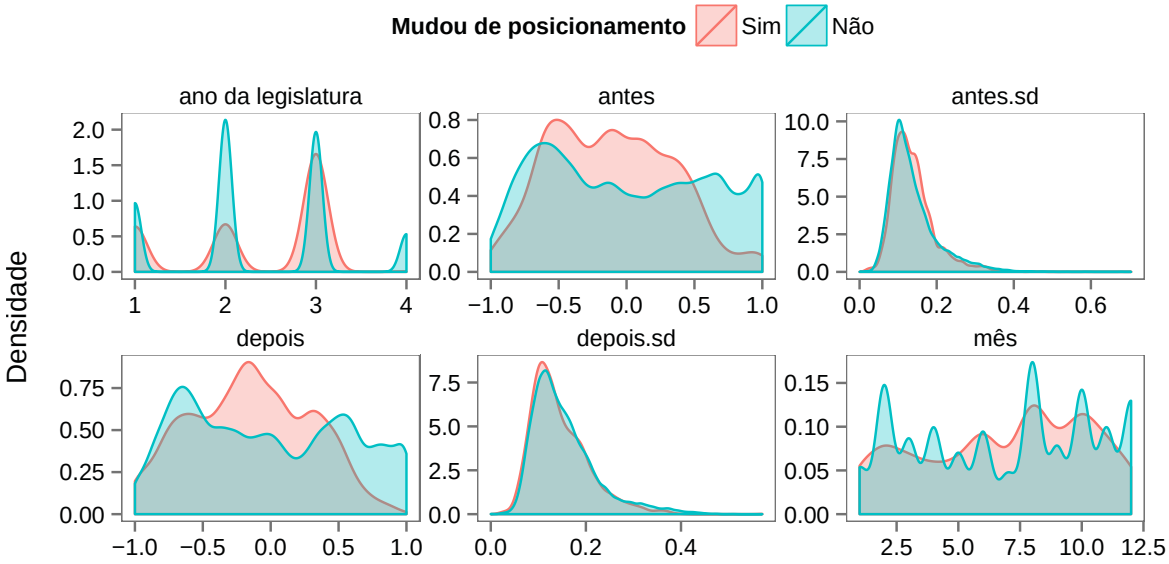


Figura 4.9: Densidade das variáveis independentes dos parlamentares em períodos que mudaram ou não de posicionamento.

evitar esse problema, é necessário testá-los em um conjunto de dados diferente do que foram treinados, o que gera a questão sobre como dividir os dados (KUHN; JOHNSON, 2013).

A forma mais simples é dividir aleatoriamente o conjunto de dados em dois subconjuntos, um para treino e outro para teste. Divisões aleatórias partem do princípio de que o padrão dos dados é uniforme; logo, não há diferença na escolha. Visto que estamos trabalhando com um período longo, inclusive com uma alteração nas regras para mudança de partido, essa é uma suposição improvável. Como é mais importante que o modelo tenha uma boa performance nos anos mais recentes, seguimos a seguinte estratégia (KUHN; JOHNSON, 2013):

1. Treina o modelo com os dados anteriores à 54ª legislatura, chamado conjunto de treino, ajustando seus parâmetros a partir da performance em 60% dos dados da 54ª legislatura, chamado conjunto de validação;

2. Ao finalizar o ajuste dos parâmetros no passo anterior, define qual modelo teve a melhor performance e treina um novo modelo, usando esses mesmos parâmetros, nos conjuntos de treino e validação juntos. Isso vai permitir ao modelo aprender os padrões da 54ª legislatura;

3. Define o ponto de corte para classificação, usando metade dos 40% restantes dos dados da 54ª legislatura, que não foram usados para validação;

4. Usa o restante dos dados da 54ª legislatura, chamado conjunto de teste, para estimar a performance final do modelo.

Assim, os 78.576 dados serão separados em 80.67% (63.386) para treino, 11.77% (9.251) para validação, 3.68% (2.894) para definir o ponto de corte e 3.88% (3.045) para teste. Eles possuem, respectivamente, 4.19%, 3.22%, 2.7% e 3.15% de mudanças de posicionamento. Aqui percebemos outro problema: as classes dos dados são desbalanceadas, com 96.02% de todas as observações sendo de parlamentares que não mudaram de posicionamento no período. Isso pode fazer com que, ao treinar os modelos nesses dados, eles priorizem a detecção da classe da maioria à revelia da minoria. Por exemplo, um modelo conseguiria 96.02% de acurácia simplesmente classificando todas observações como não sendo de mudanças de posicionamento. Para diminuir o impacto desse desbalanceamento, alguns modelos permitem a definição do custo para cada tipo de erro (falso positivo e falso negativo). Entretanto, como usaremos modelos que não permitem isso, preferi usar uma técnica mais simples: o *downsampling*.

No *downsampling*, todas as observações da categoria mais rara (mudanças de posicionamento) são selecionadas, juntamente com uma amostragem sem repetição, de tamanho similar, da categoria mais comum. Essa amostra deve manter os padrões das variáveis independentes do conjunto original. Por definição, o *downsampling* altera a distribuição real das categorias; por isso, ela só deve ser usada com os dados de treino. Caso fosse usada com os de validação ou teste, as estimativas de performance do modelo estariam enviesadas, não refletindo sua performance em dados reais.

Ao final, das 63.386 observações no conjunto de treino, 8.73% (5.531) foram usadas, sendo 48.06% de mudanças de posicionamento. Os conjuntos de validação e teste não foram modificados, mantendo-se com 9.251 e 3.045 observações, sendo, respectivamente, 3.22% e 3.15% de mudanças de posicionamento.

Natureza	Sigla	Nome
Linear	GLM	Modelo Linear Generalizado (GLM)
Não-linear	SVM Radial	SVM com kernel radial
Não-linear	RF	*Random Forest* (RF)
Não-linear	GBM	*Stochastic Gradient Boosting* (GBM)
Não-linear	C5.0	C5.0
Não-linear	NNET	Rede neural (NNET)

Tabela 4.4: Tipos de modelos preditivos testados.

4.4.4 Modelos

Wolpert (1996) afirma em seu teorema "No Free Lunch"[5] que não há uma técnica única de modelagem que seja melhor em todos casos. Por isso, Kuhn e Johnson (2013) sugerem testar diversos tipos de modelos antes de definir o que focar. Nesta seção, usarei 6 tipos: GLM, SVM com kernel radial, RF, GBM, C5.0 e NNET (Tabela 4.4). Cada um deles será treinado e validado nos conjuntos de dados descritos na Seção 4.4.3, usando todas as permutações dos parâmetros definidos no apêndice B. Os preditores utilizados serão as variáveis independentes, definidas em 4.4.2, e suas interações[6]. O critério de escolha dos modelos é a área sob a curva ROC (AUC).

A Tabela 4.5 mostra o conjunto de parâmetros de cada modelo que obteve o maior AUC no conjunto de validação, juntamente com seus intervalos de confiança em 95%[7]. O que apresentou o maior AUC foi o C5.0, com 0,97 (0,96, 0,98), contra 0,86 (0,84, 0,87) do segundo lugar, o de redes neurais. Na Figura 4.10 é possível visualizar mais facilmente a diferença do AUC entre os modelos.

Encontrados o tipo de modelo e seus parâmetros com melhor performance, treinaremos o modelo final com um novo conjunto de dados composto pela união dos conjuntos de treino

[5] Em tradução livre, "não existe almoço grátis".

[6] As interações são definidas como a multiplicação entre cada permutação das variáveis. Isso permite ao modelo usar preditores mais complexos como "fevereiro do terceiro ano da legislatura".

[7] O intervalo de confiança foi calculado usando 2.000 amostragens com reposição do conjunto de validação, mantendo as proporções das classes. Essa técnica é chamada de *bootstraping* estratificado. O número de amostragens foi baseado nos trabalhos de Carpenter e Bithell (2000) e Robin et al. (2011).

Modelo	Parâmetros		AUC		
	Nome	Valor	Mínimo	Mediano	Máximo
C5.0	Tipo	tree	0,96	0,97	0,98
	Winnow	FALSE			
	Iterações boosting	30			
NNET	Unidades ocultas	1	0,84	0,86	0,87
	Decaimento	0,1			
GBM	Encolhimento	0,1	0,82	0,83	0,85
	Profundidade máxima	4			
	Tamanho mínimo das folhas	10			
	Iterações boosting	50			
RF	Preditores aleatórios	2	0,74	0,77	0,8
SVM Radial	Custo	0,5	0,72	0,75	0,78
	Sigma	0,14			
GLM	—	—	0,38	0,41	0,43

Tabela 4.5: Lista de parâmetros dos modelos com sua respectiva área sob a curva ROC (AUC) no conjunto de validação.

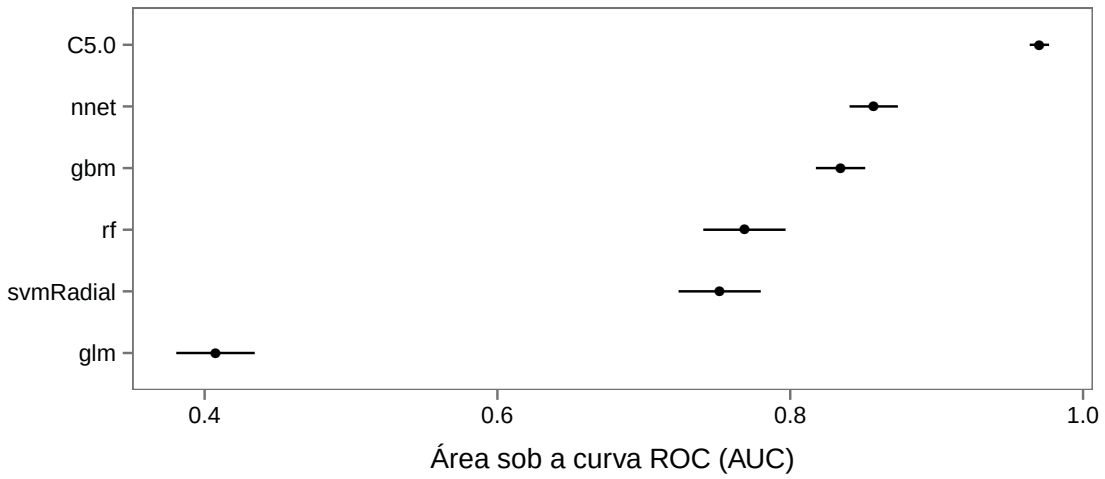


Figura 4.10: Área sob a curva ROC (AUC) dos modelos no conjunto de validação.

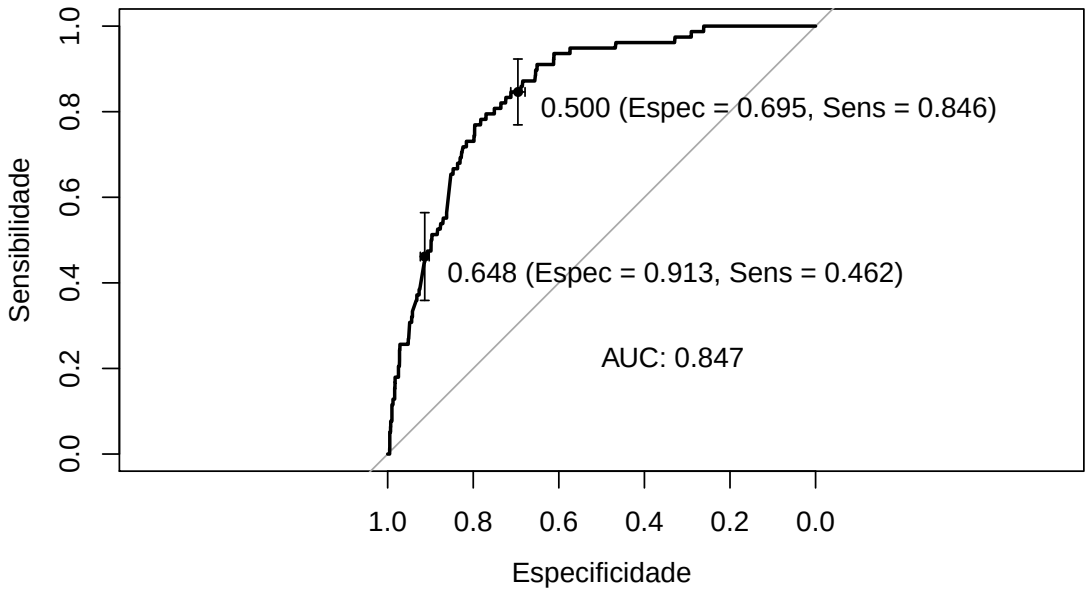


Figura 4.11: Curva ROC do modelo final no conjunto de escolha do ponto de corte.

e validação. Como o de validação não é balanceado, tendo apenas 3.22% (298) mudanças de posicionamento, ele precisa ser *downsampled*. O conjunto final terá 6.314 observações, sendo 46.82% (2.956) relativas a mudanças de posicionamento e 53.18% (3.358) não. O último passo é definir o ponto de corte para a classificação das categorias.

A maioria das ferramentas de aprendizagem de máquina define como ponto de corte padrão 0,5 (ver Seção 2.1.2). Entretanto, especialmente em casos onde as categorias não estão balanceadas ou onde o custo de um falso positivo é diferente de um falso negativo, é útil analisar pontos de corte alternativos. Para este modelo, meu objetivo é minimizar o número de falsos positivos[8], dado que 96.02% das observações são de parlamentares que não mudaram de posicionamento. Considere o seguinte exemplo:

Suponha que em um período estimamos os pontos ideais de 500 parlamentares, sendo que 95% (475) não mudaram de posicionamento, e 5% (25) mudaram. Caso o modelo tenha uma especificidade de 80%, então 20% (95) dos parlamentares que não mudaram de posicionamento serão classificados erroneamente. Mesmo que esse modelo tenha 100% de sensibilidade e classifique corretamente todos os 25 parlamentares que mudaram de posicionamento (o que é improvável), eles estarão em meio a um conjunto de 120 pessoas, o que

[8]Parlamentares que não mudaram de posicionamento, sendo classificados como se o tivessem feito.

pode inviabilizar a análise individual.

A Figura 4.11 mostra a curva ROC do modelo no conjunto de dados de definição do ponto de corte com dois pontos em destaque: o padrão 0,5 e o escolhido 0,65. Considerando o corte em 0,5, no exemplo acima 166 parlamentares seriam classificados como tendo mudado de posicionamento, sendo 87.35% (145) falsos positivos e 4 falsos negativos. Já com o corte em 0,65, 53 seriam classificados como tendo mudado de posicionamento, sendo 77.36% (41) falsos positivos e 13 falsos negativos. Em outras palavras, ao mudar o ponto de corte de 0,5 para 0,65, estamos trocando 104 falsos positivos a menos por 9 falsos negativos a mais. No apêndice C, há uma tabela com os pontos de corte, e suas respectivas sensibilidades e especificidades. Eles foram escolhidos buscando os máximos locais da curva ROC

O modelo final alcançou uma AUC 0,88 (0,86, 0,91), como visto na Figura 4.12. Considerando o ponto de corte em 0,65, o modelo possui Kappa 0,21 e 52.08% das mudanças (sensibilidade) e 91.45% das não-mudanças de posicionamento (especificidade) classificadas corretamente. A Tabela 4.6 mostra a matriz de confusão nos dados de teste. De acordo com a classificação de Hosmer Jr., Lemeshow e Sturdivant (2013) descrita na Tabela 2.2, o modelo possui um poder de discriminação excelente.

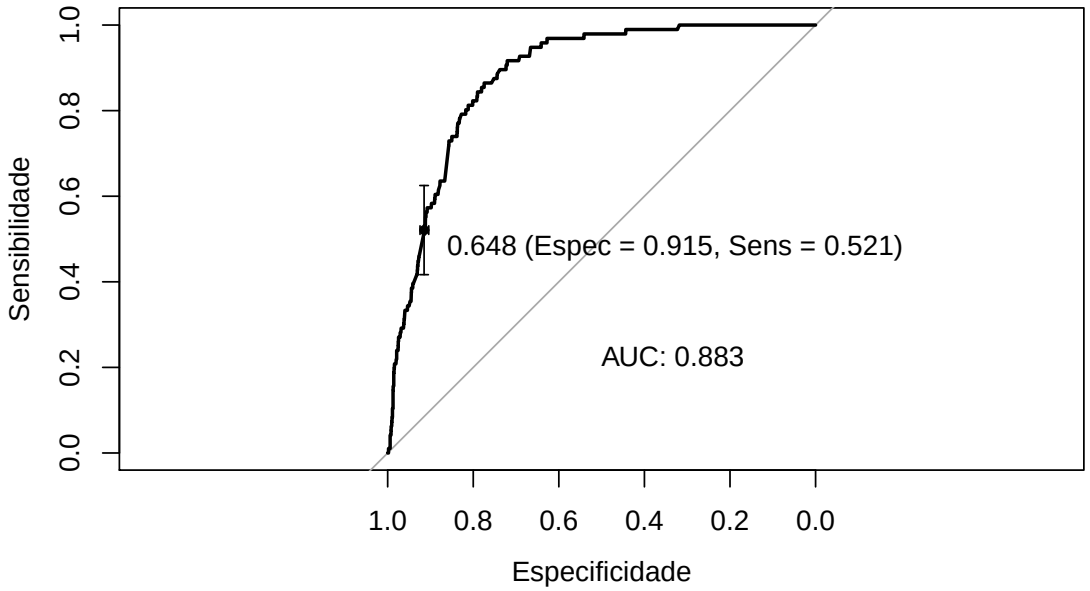


Figura 4.12: Curva ROC do modelo final nos dados de teste.

Previsto	Observado	
	Sim	Não
Sim	50	252
Não	46	2.697

Tabela 4.6: Matriz de confusão do modelo final no conjunto de dados de teste com ponto de corte em 0,65.

4.5 Considerações finais

Na primeira parte desta seção, foram apresentadas as etapas seguidas para coleta e preparação dos dados, descrevendo os parâmetros usados na estimação dos pontos ideais e a definição do "raio de influência" das mudanças de posicionamento. Após isso, os dados foram analisados, identificando suas características e aspectos temporais.

Ao final, o processo de modelagem foi apresentado, com a definição das variáveis dependente e independentes e a divisão da base de dados. Diante disso, seis tipos de modelos preditivos foram testados: GLM, SVM com kernel radial, RF, GBM, C5.0 e NNET. Com base na área sob a curva ROC (AUC), o modelo que usou C5.0 foi escolhido. Após a definição do ponto de corte, ele alcançou um AUC 0,88 no conjunto de teste, tendo um poder de discriminação excelente de acordo com a classificação de Hosmer Jr., Lemeshow e Sturdivant (2013).

Na próxima seção, apresentarei as conclusões da pesquisa, descrevendo os resultados obtidos, suas limitações e possíveis trabalhos futuros.

Capítulo 5

Conclusão

Este trabalho propôs uma solução para o problema, causado pelo volume de dados, da dificuldade em acompanhar mudanças no posicionamento dos deputados federais. Com o modelo desenvolvido, é possível descobrir quais dos parlamentares têm maior chance de ter mudado de posicionamento no período, se tornando governo ou oposição. Assim, cidadãos, jornalistas e cientistas políticos podem otimizar o uso de seu tempo.

O modelo final, que usou o método C5.0, apresentou resultados excelentes ao ser validado em um subconjunto dos dados da 54ª legislatura (Seção 4.4.4), alcançando uma área sob a curva ROC (AUC) de 0,88. Com o ponto de corte em 0,65, definido a partir da análise da performance do modelo em um conjunto de dados distinto, foram identificados corretamente 52.08% dos parlamentares que mudaram de posicionamento e 91.45% dos que não mudaram. Um erro neste trabalho significa que, de acordo com o modelo, o deputado se comportou como alguém que mudou de posicionamento, mesmo não tendo mudado até aquele momento, o que pode ser também uma informação relevante.

Na Seção 1.2, foram definidas duas perguntas de pesquisa:

1. *É possível detectar a mudança de posicionamento de um deputado federal, com ele entrando ou saindo da coalizão governamental, a partir de uma mudança no seu padrão de votação?*

2. *Os deputados federais mudam seu padrão de votação antes de mudarem de posicio-*

namento?

Com a performance alcançada pelo modelo, mostramos que é possível detectar mudanças de posicionamento a partir de mudanças de comportamento. Apesar da segunda pergunta não ter sido testada diretamente, considero essa performance como sendo um forte indicador que os deputados federais mudam de comportamento antes de mudarem de posicionamento.

Foi tomado um cuidado redobrado em tornar os resultados dessa pesquisa facilmente replicáveis com, além da própria dissertação, todos os dados e códigos-fonte dos programas desenvolvidos estando disponíveis na Internet. Eles foram licenciados livremente, permitindo que sejam usados de qualquer forma, inclusive comercial, com a única exigência sendo a citação da fonte. Assim, espero que esses resultados sejam usados na melhoria ou criação de novas ferramentas de monitoramento legislativo.

5.1 Contribuições

A principal contribuição deste trabalho é a metodologia criada para o desenvolvimento de um modelo que detecte as mudanças de posicionamento de parlamentares. Além de propor uma divisão de dados que tenta resolver problemas no treinamento do modelo em dados especialmente desafiadores, por serem temporais e desbalanceados, descrevo problemas como a definição dos períodos de análise e o "raio de influência" das mudanças de posicionamento, que não foram definidos em nenhum dos trabalhos encontrados na revisão bibliográfica.

O modelo desenvolvido e os códigos-fonte dos programas escritos para gerá-lo[1] são também contribuições interessantes, especialmente por eles serem licenciados livremente, permitindo reuso e modificação inclusive com fins comerciais.

Para extrair os dados das votações, foi desenvolvido um programa que, juntamente com os dados em si, também foi disponibilizado livremente na Internet[2].

Por fim, na análise dos dados na Seção 4.3.2 foram encontrados padrões temporais nas mudanças de posicionamento. Esses padrões foram encontrados somente nos parlamentares

[1] Disponível em `https://github.com/vitorbaptista/dissertacao-mestrado`

[2] Disponível em `https://github.com/vitorbaptista/dados-abertos-camara.gov.br`

que mudaram de posicionamento através da migração partidária, e não pelos parlamentares cujo próprio partido entrou ou saiu da coalizão. Isso reforça o que já é sabido na ciência política: migrações partidárias possuem aspectos temporais.

5.2 Limitações

Alguns problemas na modelagem não puderam ser solucionados por exigirem uma análise mais aprofundada e não serem o foco deste trabalho. Eles serão descritos nesta seção.

5.2.1 Períodos de análise

Na Seção 4.2.1.4, defini arbitrariamente os períodos de análise como sendo de 12 meses divididos em dois subperíodos de mesma duração. Uma forma melhor de defini-los seria buscando o menor período que contenha parlamentares e votações suficientes para, de acordo com os critérios definidos, estimar os pontos ideais.

5.2.2 Critérios de inclusão de parlamentares e votações

Neste trabalho, segui os critérios de inclusão padrão do W-NOMINATE, definido como ao menos 20 votações, cuja maioria foi responsável por, no máximo, 97,5% dos votos. Manter os mesmos critérios é importante em análises para permitir que os resultados sejam comparáveis com outros trabalhos. Como o objetivo do modelo não é a análise, é possível que critérios mais (ou menos) restritivos possam aumentar sua performance. Uma forma de definir esses critérios é comparar a performance de modelos treinados com os pontos ideais calculados usando critérios distintos.

5.2.3 Raio de influência da mudança de posicionamento

Uma das hipóteses básicas deste trabalho é que os parlamentares mudam de comportamento antes de mudarem oficialmente de posicionamento, permitindo assim a previsão de um a

partir do outro. Baseado nisso, foi preciso definir em quanto tempo, antes de oficializada a mudança de posicionamento, o parlamentar mudava de comportamento. Neste trabalho, esse período foi definido arbitrariamente como sendo 6 meses.

Uma forma de encontrar o "raio de influência" real é buscando o dia em que os pontos ideais nos períodos antes e depois dele estão mais distantes. É provável que esse valor varie bastante, dependendo do parlamentar e do momento, o que seria um complicador. Nesse caso, poder-se-ia buscar o valor mais comum.

5.2.4 Número de repetições para estimar o erro

Para estimar o erro dos pontos ideais, Poole (2005) sugere repetir seu cálculo 101 vezes. Como explicado na Seção 4.2.1.5, esse número se provou impraticável com o volume de dados e capacidade de processamento disponível. Por isso, neste trabalho o erro foi calculado usando 10 repetições.

Esse número foi arbitrário. É possível que a performance do modelo aumente com mais repetições. Para descobrir isso, seria preciso gerar as estimativas de erro com diversos números de repetições (inclusive zero) e encontrar o que gera um modelo com maior performance.

5.3 Trabalhos futuros

Pela pesquisa bibliográfica realizada, o uso de técnicas da ciência de dados para detectar mudanças de posicionamento dos parlamentares é uma área relativamente inexplorada. Dessa forma, existem diversos trabalhos possíveis.

Além da resolução das limitações apresentadas na Seção 5.2, o processo de modelagem descrito nesta dissertação poderia ser replicado em outra casa legislativa, como o Senado Federal ou casas de outras esferas. Inclusive, é interessante testar se os indicadores de mudanças de posicionamento são parecidos em casas distintas, analisando a performance de um modelo treinado com os dados de uma casa nos dados de outra. É possível que um modelo único treinado com os dados de diversas casas legislativas apresente uma performance

superior do que um treinado somente com os dados de cada casa.

O W-NOMINATE foi escolhido por ser um dos métodos mais usados e existirem implementações gratuitas disponíveis. Entretanto, é possível que o uso de outro método gere melhores resultados. Podem ser gerados modelos que usem mais de um método ao mesmo tempo, deixando a tarefa de descobrir qual é o mais significativo para o próprio modelo.

Neste trabalho, focamos no comportamento do parlamentar individualmente, mas é simples expandir o modelo para detectar mudanças no posicionamento dos partidos. Para isso, basta analisar o partido dos deputados que têm uma maior probabilidade de ter mudado de posicionamento. Um grande percentual, sendo de um mesmo partido, pode indicar uma mudança organizada. Esse tipo de análise pode ser expandida não só para os partidos, mas também para grupos intrapartidários como as bancadas.

Apesar dos códigos-fonte de todos os programas desenvolvidos estarem disponíveis, seu uso exige conhecimentos de computação. Integrar o modelo a ferramentas como o Basômetro ou o Radar Parlamentar são trabalhos essenciais para que os benefícios deste trabalho sejam aproveitados por mais pessoas.

Bibliografia

ABRANCHES, S. H. H. D. Presidencialismo de coalizão: o dilema institucional brasileiro. *Revista de Ciências Sociais*, v. 31, n. 1, p. 5–34, 1988.

ALTMAN, D. G.; BLAND, J. M. Diagnostic tests 3: Receiver Operating Characteristic plots. *British Medical Journal*, v. 309, n. 6948, p. 188, 1994.

AMES, B. *Os entraves da democracia no Brasil*. [S.l.: s.n.], 2003.

ANDRADE, N.; MONTEIRO, J. a. A. B. *House of Cunha*. 2015. Disponível em: <http://houseofcunha.com.br>.

ARAÚJO, M. *Migração Partidária: A evolução das migrações partidárias na Câmara dos Deputados – 1979/1999*. Dissertação (Mestrado) — Universidade de São Paulo, 2000.

BAILEY, M. A. Time Comparable Preference Estimates across and Institutions for the and Presidency. *American Journal of Political Science*, v. 51, n. 3, p. 433–448, 2007. Disponível em: <http://faculty.georgetown.edu/baileyma/ajps_offprint_bailey.pdf>.

BAPTISTA, V. M. P. d. S. et al. Uma ferramenta para analisar mudanças na coesão entre parlamentares em votações nominais. In: *III Brazilian Workshop on Social Network Analysis and Mining*. Brasília: [s.n.], 2014. Disponível em: <https://www.academia.edu/15002720>.

BERNABEL, R. Does the Electoral Rule Matter for Political Polarization? The Case of Brazilian Legislative Chambers. *Brazilian Political Science Review*, São Paulo, v. 9, n. 2, p. 81–108, 2015. Disponível em: <http://ref.scielo.org/dz55w6>.

BRASIL. *Lei nº 9.504, de 30 de setembro de 1997. Estabelece normas para as eleições*. 1997. Disponível em: <https://www.planalto.gov.br/ccivil_03/leis/l9504.htm>.

BREVE, N. Direção do pt expulsa rebeldes e adverte tendências radicais. 2003. Disponível em: <http://www.cartamaior.com.br/?/Editoria/Politica/Direcao-do-PT-expulsa-rebeldes-e-adverte-tendencias-radicais/4/733>.

BROWN, C. D.; DAVIS, H. T. Receiver operating characteristics curves and related decision measures: A tutorial. *Chemometrics and Intelligent Laboratory Systems*, v. 80, n. 1, p. 24–38, 2006.

Câmara dos Deputados. *Dados Abertos – Legislativo*. 2015. Disponível em: <http://www2-.camara.leg.br/transparencia/dados-abertos/dados-abertos-legislativo>.

CARNEIRO, A. C. d. S.; SANTOS, L. C. A. dos; NETO, M. G. d. N. *Curso de Regimento Interno*. 2. ed. Brasília: Edições Câmara, 2013. 466 p. ISBN 978-85-402-0178-1.

CARPENTER, J.; BITHELL, J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics in medicine*, v. 19, n. 9, p. 1141–1164, 2000. ISSN 1097-0258.

CHEIBUB, J. A.; FIGUEIREDO, A.; LIMONGI, F. Partidos políticos e governadores como determinantes do comportamento legislativo na câmara dos deputados, 1988-2006. *DADOS - Revista de Ciências Sociais*, v. 52, n. 2, p. 263–299, 2009. Disponível em: <http://www.scielo.br/pdf/dados/v52n2/v52n2a01.pdf>.

CLINTON, J.; JACKMAN, S.; RIVERS, D. The statistical analysis of roll call data. *American Political Science ...*, v. 98, n. 2, p. 355–370, 2004. Disponível em: <http://journals.cambridge.org/abstract_S0003055404001194>.

COHEN, J. A coefficient of agreement for nominal data. *Educational and Psychological Measurement*, v. 20, n. 1, p. 37–46, 1960.

CONOVER, M. D. et al. Predicting the Political Alignment of Twitter Users. In: *IEEE Third Inernational Conference on Social Computing (SocialCom)*. [S.l.: s.n.], 2011. p. 192–199.

CONWAY, D. *The Data Science Venn Diagram*. 2013. Disponível em: <http://drewconway-.com/zia/2013/3/26/the-data-science-venn-diagram>.

DANTAS, H.; TOLEDO, J. R. de; TEIXEIRA, M. A. C. *Análise política & jornalismo de dados*. Rio de Janeiro: Editora FGV, 2014. 226 p. ISBN 9788522515011.

DESPOSATO, S. W. The Impact of Party-Switching on Legislative Behavior in Brazil. 2005. Disponível em: <http://swd.ucsd.edu/ps2.pdf>.

ESTADÃO. *Basômetro*. São Paulo: [s.n.], 2012. Disponível em: <http://estadaodados.com-/basometro/>.

FAWCETT, T. An introduction to ROC analysis. *Pattern recognition letters*, v. 27, n. 8, p. 861–874, 2006.

FIGUEIREDO, A. C. Government Coalitions in Brazilian democracy. *Brazilian Political Science Review*, v. 1, n. 2, p. 182–216, 2007. Disponível em: <http://www.bpsr.org.br-/index.php/bpsr/article/view/52>.

FIGUEIREDO, A. C.; LIMONGI, F. *Executivo e Legislativo na nova ordem constitucional*. 2. ed. Rio de Janeiro: [s.n.], 2001. 232 p. ISSN 8522502919. ISBN 8522502919.

Fiscal Note. *Prophecy*. 2015. Disponível em: <https://www.fiscalnote.com>.

FREITAS, A. M. D. *Migração partidária na câmara dos deputados*. Dissertação (Mestrado) — Universidade de São Paulo, 2008. Disponível em: <http://www.teses.usp.br/teses-/disponiveis/8/8131/tde-11112009-151004/pt-br.php>.

FREITAS, A. M. de; IZUMI, M. Y.; MEDEIROS, D. B. O Congresso Nacional em duas dimensões: estimando pontos ideais de deputados e senadores (1988-2010). In: *8º Encontro da ABCP*. [s.n.], 2012. Disponível em: <http://www.academia.edu/4176280-/O_Congresso_Nacional_em_duas_dimens\~oes_estimando_pontos_ideais_de_Deputados_e_Senadores_2010>.

GEISSER, S. *Predictive Inference: An Introduction*. [S.l.: s.n.], 1993. 240 p.

GERRISH, S. M.; BLEI, D. M. Predicting legislative roll calls from text. In: *Proceedings of the 28th International Conference on Machine Learning*. Bellevue, WA: [s.n.], 2011. Disponível em: <http://machinelearning.wustl.edu/mlpapers/paper_files-/ICML2011Gerrish_333.pdf>.

GOLDBLATT, D.; O'NEIL, T. How a Bill Becomes a Law - Predicting Votes from Legislation Text. 2012. Disponível em: <http://cs229.stanford.edu/projects2012.html>.

HOFFMAN, P. et al. *Scrapy*. 2014. Disponível em: <http://python.org>.

HOSMER JR., D. W.; LEMESHOW, S.; STURDIVANT, R. X. *Applied logistic regression*. 3. ed. [S.l.]: Wiley, 2013. 448 p.

IZUMI, M. Y. *Os determinantes do comportamento parlamentar no Senado Brasileiro (1989-2010)*. 86 p. Dissertação (Master) — Universidade de São Paulo, 2013. Disponível em: <http://www.teses.usp.br/teses/disponiveis/8/8131/tde-19022014-124813/pt-br.php>.

JACKMAN, S. Estimation and Inference via Bayesian Simulation: An Introduction to Markov Chain Monte Carlo. *American Journal of Political Science*, v. 44, n. 2, p. 375–404, 2000. Disponível em: <http://www.jstor.org/stable/2669318>.

KUHN, M.; JOHNSON, K. *Applied Predictive Modeling*. [s.n.], 2013. 1–595 p. ISBN 978-1-4614-6848-6. Disponível em: <http://link.springer.com/10.1007/978-1-4614-6849-3>.

LAMOUNIER, B. A Democracia Brasileira de 1985 à Década de 90: A Síndrome da Paralisia Hiperativa. In: OLYMPIO, J. (Ed.). *Governabilidade, sistema político e violência urbana*. 1. ed. [S.l.: s.n.], 1994. ISBN 9.788503005296E12.

LEONI, E. Ideologia, democracia e comportamento parlamentar: a Câmara dos Deputados (1991-1998). *Dados*, v. 45, n. 3, p. 361–386, 2002. ISSN 0011-5258. Disponível em: <http://www.scielo.br/pdf/dados/v45n3/a02v45n3>.

LEVY, S. The AI revolution is on. *Wired*, v. 19, n. 1, p. 86–97, dez. 2010. Disponível em: <http://www.wired.com/2010/12/ff_ai_essay_airevolution/>.

LIMONGI, F.; FIGUEIREDO, A. C. Partidos políticos na câmara dos deputados: 1989-1994. *DADOS - Revista de Ciências Sociais*, v. 38, n. 3, p. 497–525, 1995. Disponível em: <http://www.fflch.usp.br/dcp/assets/docs/Limongi-/Partidos_Politicos_na_Camara_dos_Deputados_1989-1994.pdf>.

MAINWARING, S. P. *Sistemas partidários em novas democracias: o caso do Brasil*. 1. ed. Porto Alegre: [s.n.], 2001.

MCCARTY, N. M. Measuring Legislative Preferences. In: *The Oxford Handbook of the American Congress*. [s.n.], 2011. ISBN 9780199559947. Disponível em: <http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199559947.001.0001-/oxfordhb-9780199559947-e-4>.

MCFADDEN, D. L. Quantal Choice Analysis: A Survey. *Annals of Economic and Social Measurement*, v. 5, n. 4, p. 363–390, 1976. Disponível em: <http://www.nber.org/chapters-/c10488.pdf>.

MELO, C. R. *Retirando as cadeiras do lugar: migração partidária na Câmara dos Deputados (1985-2002)*. Belo Horizonte: Editora UFMG, 2004. 212 p. ISBN 85-7041-433-1.

MÜLLER, W. C.; STROM, K. *Coalition Governments in Western Europe*. Londres: [s.n.], 2000.

NG, A. *Machine Learning*. 2012. Disponível em: <https://www.coursera.org/learn-/machine-learning>.

POOLE, K. et al. Scaling Roll Call Votes with wnominate in R. *Journal of Statistical Software*, v. 42, n. 14, 2011. Disponível em: <http://www.jstatsoft.org/v42/i14/paper>.

POOLE, K. T. Nonparametric unfolding of binary choice data. *Political Analysis*, v. 8, n. 3, p. 211–237, 2000. ISSN 1047-1987. Disponível em: <http://polmeth.wustl.edu/analysis-/vol/8/PA83-211-237.pdf>.

POOLE, K. T. *Spatial models of parliamentary voting*. [S.l.]: Cambridge University Press, 2005.

POOLE, K. T.; ROSENTHAL, H. A Spatial Model for Legislative Roll Call Analysis. *American Journal of Political Science*, v. 29, n. 2, p. 357–384, 1985. Disponível em: <http://www.jstor.org/stable/2111172>.

PORTO, F. A.; ZIVIANI, A. Ciência de Dados. In: *3º Seminário Grandes Desafios da Computação no Brasil Fase 2*. Rio de Janeiro: [s.n.], 2014.

POWER, T. J. Optimism, pessimism, and coalitional presidentialism: Debating the institutional design of Brazilian democracy. *Bulletin of Latin American Research*, v. 29, n. 1, p. 18–33, 2010. ISSN 02613050.

Python Software Foundation. *Python*. 2014. Disponível em: <http://python.org>.

QUINN, K. M. et al. An Automated Method of Topic-Coding Legislative Speech Over Time with Application to the 105th-108th U.S. Senate. 2006. Disponível em: <http://www.allacademic.com/meta/p150986_index.html>.

RICE, S. A. *Farmers and workers in american politics*. Tese (Doutorado) — Columbia University, 1924.

ROBIN, X. et al. proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, v. 12, p. 77, 2011.

SANTOS, F. G. M. *O poder legislativo no presidencialismo de coalizão*. [S.l.: s.n.], 2003. 251 p.

SHOR, B.; BERRY, C. R.; MCCARTY, N. M. A Bridge to Somewhere: Mapping State and Congressional Ideology on a Cross-Institutional Common Space. *Legislative Studies Quarterly*, v. 35, n. 3, p. 1–32, 2010. ISSN 1939-9162.

SOMASUNDARAN, S.; WIEBE, J. Recognizing stances in ideological on-line debates. In: *Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*. [s.n.], 2010. p. 116–124. Disponível em: <http://dl.acm.org/citation.cfm?id=1860645>.

STANTON, J. *An Introduction to Data Science*. [s.n.], 2012. 157 p. ISBN 9781491901427. Disponível em: <http://jsresearch.net/groups/teachdatascience/wiki/welcome/attachments-/72f24/DataScienceBook1_1.pdf\$backslash$npapers2://publication/uuid/99B2E09F-00FE-448F-8E88-89102110B293>.

THOMAS, M.; PANG, B.; LEE, L. Get out the vote: determining support or opposition from congressional floor-debate transcripts. In: *Conference on Empirical Methods in Natural Language Processing*. [s.n.], 2006. p. 129–137. Disponível em: <http://dl.acm.org/citation.cfm?id=1610122>.

TRENTO, S.; LEITE, L. Parlamentares no Radar Parlamentar. 2013. Disponível em: <https://github.com/radar-parlamentar/radar/raw-/8d54db1df3830144892b93e799f86b5c46c00576/doc/radar_parlamentar.pdf>.

Tribunal Superior Eleitoral. Resolução nº 22.610. Brasília, DF, out. 2007. Disponível em: <http://www.tse.jus.br/legislacao/codigo-eleitoral/normas-editadas-pelo-tse/resolucao-nb0-22.610-de-25-de-outubro-de-2007-brasilia-2013-df>.

VICENTE, P. C. *Prevalência da câmara dos deputados no processo legislativo bicameral: a lei de improbidade administrativa*. Brasília: Biblioteca Digital da Câmara dos Deputados, 2012. 128 p. Disponível em: <http://bd.camara.gov.br/bd/bitstream/handle/bdcamara/8708-/prevalencia_camara_vicente.pdf?sequence=1>.

WANG, J.; VARSHNEY, K. R.; MOJSILOVI, A. Legislative Prediction via Random Walks over a Heterogeneous Graph. In: *SDM*. [S.l.: s.n.], 2012. p. 1095–1106. ISBN 9781611972320.

WOLPERT, D. H. The lack of a priori distinctions between learning algorithms. *Neural computation*, v. 8, n. 7, p. 1341–1390, 1996.

YANO, T.; SMITH, N. A.; WILKERSON, J. D. Textual predictors of bill survival in congressional committees. In: *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. [S.l.: s.n.], 2012. p. 793–802.

YU, B.; KAUFMANN, S.; DIERMEIER, D. Classifying Party Affiliation from Political Speech. *Journal of Information Technology & Politics*, v. 5, n. 1, p. 33–48, 2008. Disponível em: <http://www.tandfonline.com/doi/abs/10.1080/19331680802149608>.

ZUCCO, C. Ideology or What? Legislative Behavior in Multiparty Presidential Settings. *The Journal of Politics*, v. 71, n. 03, p. 1076, 2009. ISSN 0022-3816. Disponível em: <http://www.jstor.org/stable/10.1017/S0022381609090896>.

ZUMEL, N.; MOUNT, J. *Practical Data Science with R*. Manning Publications Co., 2014. 1–417 p. ISBN 9781617291562. Disponível em: <http://dl.acm.org/citation-.cfm?id=2614429>.

Apêndice A

Composição das coalizões governamentais

Legislatura	Governo	Data inicial	Composição
50ª	FHC I	1995/01/01	PFL, PMDB, PSDB e PTB
		1996/04/26	PFL, PMDB, PPB, PSDB e PTB
51ª	FHC II	1999/01/01	PFL, PMDB, PPB e PSDB
		2002/03/06	PMDB, PPB e PSDB
52ª	Lula I	2003/01/01	PCdoB, PDT, PL, PPS, PSB, PT, PTB e PV
		2004/01/23	PCdoB, PL, PMDB, PPS, PSB, PT, PTB e PV
		2005/02/01	PCdoB, PL, PMDB, PSB, PT, PTB e PV
		2005/05/20	PCdoB, PL, PMDB, PSB, PT e PTB
		2005/07/23	PCdoB, PL, PMDB, PP, PSB, PT e PTB
53ª	Lula II	2007/01/01	PCdoB, PMDB, PP, PR, PRB, PSB e PT
		2007/04/02	PCdoB, PDT, PMDB, PP, PR, PRB, PSB, PT e PTB
		2009/09/28	PCdoB, PDT, PMDB, PP, PR, PRB, PSB e PT
54ª	Dilma I	2011/01/01	PCdoB, PDT, PMDB, PP, PR, PSB e PT
		2012/03/02	PCdoB, PDT, PMDB, PP, PR, PRB, PSB e PT
		2013/10/03	PCdoB, PDT, PMDB, PP, PR, PRB e PT

Tabela A.1: Lista das coalizões que ocorreram durante o período de estudo e suas respectivas composições.

Apêndice B

Parâmetros dos modelos

Modelo	Parâmetro	Valores
C5.0	Tipo	rules, tree
	Winnow	FALSE, TRUE
	Iterações boosting	1, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100
NNET	Unidades Ocultas	1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39
	Decaimento	0, 1e-04, 0,00015, 0,00022, 0,00032, 0,00046, 0,00068, 0,001, 0,0015, 0,0022, 0,0032, 0,0046, 0,0068, 0,01, 0,015, 0,022, 0,032, 0,046, 0,068, 0,1
GBM	Encolhimento	0,1
	Profundidade máxima	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20
	Tamanho mínimo das folhas	10
	Iterações boosting	50, 100, 150, 200, 250, 300, 350, 400, 450, 500, 550, 600,

		650, 700, 750, 800, 850, 900, 950, 1.000
RF	Preditores aleatórios	2, 3, 5, 6, 9, 13, 17, 24, 33, 45, 62, 85, 117, 160, 219, 299, 410, 560, 766
SVM Radial	Custo	0,25, 0,5, 1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1.024, 2.048, 4.096, 8.192, 16.384, 32.768, 65.536, 131.072
	Sigma	0,14
GLM	—	—

Tabela B.1: Lista dos parâmetros usados no treinamento dos modelos. Em modelos com mais de um parâmetro, foram testadas todas suas permutações.

Apêndice C

Pontos de corte do modelo final

Threshold	Especificidade			Sensibilidade		
	Mínimo	Mediano	Máximo	Mínimo	Mediano	Máximo
0.079	24.47%	26.1%	27.77%	100%	100%	100%
0.108	27.31%	29.01%	30.72%	96.15%	98.72%	100%
0.154	31.11%	32.81%	34.59%	93.59%	97.44%	100%
0.307	44.82%	46.66%	48.54%	91.03%	96.15%	100%
0.402	55.5%	57.35%	59.16%	89.74%	94.87%	98.72%
0.432	59.34%	61.12%	62.96%	87.18%	93.59%	98.72%
0.433	59.41%	61.19%	63.03%	85.9%	92.31%	97.44%
0.477	63.42%	65.06%	66.94%	84.62%	91.03%	97.44%
0.481	63.71%	65.38%	67.22%	82.05%	89.74%	96.15%
0.482	63.78%	65.45%	67.29%	80.77%	88.46%	94.87%
0.494	66.69%	68.41%	70.17%	79.49%	87.18%	94.87%
0.495	67.08%	68.82%	70.56%	78.21%	85.9%	93.59%
0.523	69.53%	71.2%	72.94%	76.92%	84.62%	92.31%
0.528	70.74%	72.37%	74.04%	74.36%	83.33%	91.03%
0.53	71.84%	73.54%	75.14%	73.08%	82.05%	91.03%
0.533	73.44%	75.07%	76.63%	71.79%	80.77%	89.74%
0.558	75.46%	77.02%	78.55%	70.51%	79.49%	88.46%
0.567	76.7%	78.23%	79.69%	69.23%	78.21%	87.18%

0.573	78.16%	79.65%	81.11%	67.95%	76.92%	85.9%
0.573	78.2%	79.69%	81.14%	65.38%	75.64%	84.62%
0.574	78.34%	79.83%	81.29%	64.1%	74.36%	83.33%
0.583	80.18%	81.61%	82.99%	64.07%	73.08%	83.33%
0.602	81%	82.39%	83.77%	61.54%	71.79%	82.05%
0.603	81.36%	82.71%	84.06%	60.26%	70.51%	80.77%
0.605	81.68%	83.06%	84.34%	58.97%	69.23%	79.49%
0.607	82.32%	83.66%	84.98%	57.69%	67.95%	78.21%
0.61	83.35%	84.71%	85.9%	56.41%	66.67%	76.92%
0.613	83.91%	85.23%	86.47%	55.13%	65.38%	76.92%
0.613	84.91%	86.22%	87.46%	46.15%	56.41%	67.95%
0.614	85.72%	87%	88.21%	44.87%	55.13%	66.67%
0.617	86.33%	87.61%	88.78%	42.31%	53.85%	65.38%
0.617	87.07%	88.32%	89.45%	41.03%	52.56%	64.1%
0.62	88.46%	89.7%	90.7%	39.74%	51.28%	62.82%
0.621	88.67%	89.84%	90.87%	38.46%	50%	61.54%
0.625	89.6%	90.7%	91.69%	35.9%	47.44%	58.97%
0.648	90.27%	91.34%	92.33%	34.62%	46.15%	57.69%
0.649	90.41%	91.44%	92.44%	33.33%	44.87%	56.41%
0.649	91.65%	92.65%	93.57%	26.92%	38.46%	50%
0.65	92.22%	93.18%	94.07%	26.92%	37.18%	48.72%
0.651	92.37%	93.32%	94.18%	25.64%	35.9%	47.44%
0.651	93.36%	94.21%	95.03%	23.08%	33.33%	43.59%
0.652	93.54%	94.39%	95.17%	21.79%	32.05%	42.31%
0.653	94%	94.89%	95.63%	20.51%	30.77%	41.03%
0.654	94.32%	95.17%	95.92%	16.67%	26.92%	37.18%
0.657	96.45%	97.12%	97.73%	16.67%	25.64%	35.9%
0.657	96.59%	97.23%	97.8%	15.38%	24.36%	34.62%
0.657	96.91%	97.5%	98.01%	11.54%	20.51%	29.49%
0.66	97.76%	98.26%	98.72%	10.26%	17.95%	26.92%
0.662	97.8%	98.3%	98.76%	8.97%	16.67%	25.64%

0.668	97.9%	98.37%	98.79%	7.69%	15.38%	24.36%
0.687	98.33%	98.76%	99.15%	6.38%	12.82%	20.51%
0.689	98.62%	99.01%	99.36%	5.13%	11.54%	19.23%
0.699	98.97%	99.29%	99.57%	2.56%	7.69%	14.1%
0.702	99.01%	99.33%	99.61%	1.28%	6.41%	12.82%
0.725	99.22%	99.5%	99.75%	1.28%	5.13%	10.26%
0.73	99.29%	99.54%	99.79%	0%	1.28%	3.85%
0.75	99.33%	99.57%	99.82%	0%	0%	0%
0.758	99.36%	99.61%	99.82%	0%	0%	0%
0.766	99.47%	99.68%	99.86%	0%	0%	0%
0.77	99.5%	99.72%	99.89%	0%	0%	0%
0.779	99.54%	99.75%	99.93%	0%	0%	0%
0.792	99.61%	99.79%	99.93%	0%	0%	0%
0.798	99.64%	99.82%	99.96%	0%	0%	0%
0.816	99.72%	99.86%	99.96%	0%	0%	0%
0.84	99.75%	99.89%	100%	0%	0%	0%
0.887	99.82%	99.93%	100%	0%	0%	0%
0.953	99.89%	99.96%	100%	0%	0%	0%
Inf	100%	100%	100%	0%	0%	0%

Tabela C.1: Especificidade e sensibilidade em diversos pontos de corte do modelo final usando o conjunto de dados para definição do ponto de corte.

Apêndice D

Versões dos softwares utilizados

- R version 3.2.1 (2015-06-18), `x86_64-pc-linux-gnu`
- Locale: `LC_CTYPE=en_US.UTF-8`, `LC_NUMERIC=C`, `LC_TIME=pt_BR.UTF-8`, `LC_COLLATE=en_US.UTF-8`, `LC_MONETARY=pt_BR.UTF-8`, `LC_MESSAGES=en_US.UTF-8`, `LC_PAPER=pt_BR.UTF-8`, `LC_NAME=C`, `LC_ADDRESS=C`, `LC_TELEPHONE=C`, `LC_MEASUREMENT=pt_BR.UTF-8`, `LC_IDENTIFICATION=C`
- Base packages: base, datasets, graphics, grDevices, methods, parallel, splines, stats, utils
- Other packages: C50 0.1.0-24, caret 6.0-47, data.table 1.9.4, doParallel 1.0.8, foreach 1.4.2, gbm 2.1.1, ggplot2 1.0.1, ggthemes 2.2.1, iterators 1.0.7, kernlab 0.9-20, knitr 1.9, lattice 0.20-33, nnet 7.3-10, plyr 1.8.2, pROC 1.8, randomForest 4.6-10, scales 0.2.4, survival 2.38-3, tikzDevice 0.8.1
- Loaded via a namespace (and not attached): BradleyTerry2 1.0-6, brglm 0.5-9, car 2.0-25, chron 2.3-45, codetools 0.2-11, colorspace 1.2-4, digest 0.6.8, evaluate 0.5.5, filehash 2.2-2, formatR 1.2, grid 3.2.1, gtable 0.1.2, gtools 3.4.2, highr 0.5, lme4 1.1-7, MASS 7.3-43, Matrix 1.2-2, mgcv 1.8-7, minqa 1.2.4, munsell 0.4.2, nlme 3.1-121, nloptr 1.0.4, partykit 1.0-2, pbkrtest 0.4-2, proto 0.3-10,

quantreg 5.11, Rcpp 0.11.4, reshape2 1.4.1, SparseM 1.6, stringr 0.6.2, tools 3.2.1