Universidade Federal Fluminense

FACULDADE DE ECONOMIA

PEDRO CAVALCANTE OLIVEIRA

COMPUTAÇÃO DE EFEITOS MARGINAIS EM FLORESTAS ALEATÓRIAS

NITERÓI – RJ

2021

PEDRO CAVALCANTE OLIVEIRA

COMPUTAÇÃO DE EFEITOS MARGINAIS EM FLORESTAS ALEATÓRIAS

Monografia apresentada ao curso de Bacharelado em Ciências Econômicas da Universidade Federal Fluminense como requisito parcial para conclusão do curso.

Orientador:
Prof. Dr. Jesus Alexei Luizar Obregon

Coorientador:
Prof. Dr. Bruno Santiago

Niterói – RJ

2021

PEDRO CAVALCANTE OLIVEIRA

COMPUTAÇÃO DE EFEITOS MARGINAIS EM FLORESTAS ALEATÓRIAS

Monografia apresentada ao curso de Bacharelado em Ciências Econômicas da Universidade Federal Fluminense como requisito parcial para conclusão do curso.

Trabalho aprovado em 13 de Fevereiro de 2021

BANCA EXAMINADORA

Prof. Dr. Jesus Alexei Luizar Obregon
Orientador
Universidade Federal Fluminense

Prof. Dr. Bruno Santiago
Coorientador
Universidade Federal Fluminense

Prof. Dr. Luciano Vereda
Universidade Federal Fluminense

RESUMO

Este trabalho demonstra a construção partindo de primeiros princípios de modelos de floresta aleatória, os compara com modelos lineares da Econometria Clássica e expõe a problemática de recuperar os efeitos marginais. Um algoritmo de computação dos efeitos marginais a partir de uma floresta aleatória treinada, uma simulação de Monte Carlo mostrando em ambiente controlado sua aplicação e uma aplicação prática com dados de preços de imóveis são apresentados. Por fim, agendas futuras de pesquisa são esboçadas.

Palavras-chave: Florestas Aleatórias; Aprendizado de Máquina; Econometria.

ABSTRACT

This work demonstrates the construction starting from the first principles of random forest models, compares them with linear models of Classical Econometrics and exposes the problem of recovering marginal effects. An algorithm for computing the marginal effects from a trained random forest, a Monte Carlo simulation showing its application in a controlled environment and a practical application with real estate price data are presented. Finally, future research agendas are outlined.

Keywords: Random Forests; Machine Learning; Econometrics.

AGRADECIMENTOS

Tive o enorme privilégio de crescer em um lar cheio de amor e dedicação. É algo que a passagem do tempo deixa continuamente mais claro. Devo tudo à minha (grande) família e suas figuras curiosas. Luiza, Carmosa, José Carlos, Maria Helena, Joana e outros tantos nomes queridos são figuras essenciais para mim.

Aos amigos, agradeço profundamente pela presença. Leonardo(s), Mauro, Rafael, Marcelo(s), João Pedro, Maurício, Rômulo, Ana Luiza, José Victor, Lucas, Carolina, Caetano e tantos outros que fatalmente não citarei. Obrigado por tudo.

Devo muito aos colegas de profissão. Jamil Civitarese, Tiago Dantas, Filipe Russo, Maíra Franca, Armando Martins, Flavio Abdenur, Diego Cardoso foram figuras importantes, sempre dispostas a compartilhar alguns centavos de experiência. Patrick Maia, um chefe-colega que muito me ensinou. Daniel Duque, especialmente, por ter acompanhado esses anos de graduação com um cuidado muito bonito. Mais ainda foi vê-lo se tornar uma muito necessária figura pública.

Aos mestres, agradeço muito a Paulo Gusmão, Ana Urraca, Juliana Coelho, Wilson Calmon, Diogo Bravo e Carlos Gabriel Guimarães. Devo agradecimentos especiais aos professores Jesus Alexei Luizar Obregon, que não sabe disso mas foi um dos poucos motivos que me impediram de trocar de graduação em certo momento, Luciano Vereda que quase me convenceu de que macroeconomia é um campo interessante e Bruno Santiago que com um sorriso, um café e paciência aparentemente infinita me ajudou a dar os doloridos passos iniciais na matemática de gente grande.

Por fim, agradeço ao Instituto de Pesquisa Econômica Aplicada (IPEA) do Rio de Janeiro, à Escola Brasileira de Administração Pública e Privada (EBAPE/FGV) e à Análise Macro pelas oportunidades de estágio e aprendizado. Também listo a Genesis Library, que viabilizou os meus estudos em vários momentos chave.

A todos que participaram e não foram citados, agradeço enormemente.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Nos últimos 20 anos o volume e a variedade de dados produzidos e armazenados pela humanidade aumentou em algumas ordens de grandeza. Imagens, áudio, registros de viagens, redes sociais, microdados administrativos, exames médicos, dados genômicos, a lista é vasta. Acompanhando esse movimento, principalmente na indústria de tecnologia, as aplicações de Aprendizado de Máquina (*Machine Learning*) aumentaram proporcionalmente.

O campo certamente não nasceu nos corredores do Vale do Silício. As primeiras contribuições formais na área são muito anteriores à qualquer forma de indústria de computação moderna e vêm da psicologia da consciência dos anos 40-50. Trabalhos como McCulloch e Pitts (1943) e Rosenblatt (1958) introduziram as primeiras redes neurais. Ferramentas similares começam a ser abordadas por estatísticos, procurando desempenho preditivo, e cientistas da computação, procurando inteligência artificial e automatização de processos, nos anos 70 e 80. É aí que surgem os SVMs (VAPNIK; CHERVONENKIS, 1974) e Árvores de Decisão (BREIMAN et al., 1984), partes do cânone da área.

Por ser um campo altamente interdisciplinar que prosperou na indústria e cujas técnicas são aplicadas com grande variação do software que as implementam, do ponto de vista do econometrista acostumado com manuais altamente consistentes entre si há pouca sistematização e consenso. Existem referências importantes como Friedman, Hastie e Tibshirani (2001), mas a uniformidade de aplicação dessas técnicas é muito menor do que as suas primas vindas da estatística clássica. Isso reflete um cisma que Breiman (2001) coloca em 'duas culturas da modelagem estatística'.

A primeira, que o autor chama de 'modelagem baseada nos dados' assume que os dados coletados são gerados a partir de um processo estocástico que pode ser estimado. As noções de sucesso de modelagem, nessa abordagem, são cumprir uma bateria de testes estatísticos e boas leituras em indicadores de qualidade de ajuste. A segunda, 'modelagem algorítmica', considera o processo estocástico verdadeiro a ser aproximado potencialmente complexo demais e aborda a questão com caixas-pretas a serem estimadas com processos de decisão que em teoria seriam aproximações de processos cognitivos: encontrar hiperplanos separadores, partições recursivas de conjuntos, agrupamento por proximidade, decomposição em componentes principais, compor afirmações verdadeiro-falso, agrupamento hierárquico, etc. Nessa concepção, a validade de um modelo é determinada pela sua capacidade preditiva.

Uma caixa-preta, no entanto, talvez não seja de tanto interesse ao econometrista aplicado que está sim preocupado com alguma forma de inferência. Em particular, distinguir estatisticamente efeitos de tratamento de zero. Existem métodos de interpretação de modelos (RIBEIRO; SINGH; GUESTRIN, 2016), mas pouca sistematização, em particular na literatura em economia

aplicada e econometria.

O tema é relevante em dois sentidos. Do ponto de vista do aprendizado de máquina, porque encaixa em uma agenda maior de pesquisa em *interpretabilidade* de Machine Learning. O núcleo duro reduzido da disciplina abre espaço para uma série de más práticas disseminadas no uso dessas técnicas (FLACH, 2019). Avaliar efeitos marginais pode ser usado como uma forma de validação qualitativa também. Relações com sinais inversos ao esperado podem ser evidência de problemas na estimação, entendimento do problema ou preparação dos dados.

Há também, pelo mesmo motivo, dificuldade de comunicação de resultados de modelos e até mesmo responsabilização civil-criminal quanto às consequências de seu uso em ambiente de produção, sem supervisão humana (LEPRI et al., 2018). Interpretação de modelo, em particular antes de entrega para algum ambiente de produção em que seus resultados afetarão desde experiência de uso em aplicativos de jogos à possivelmente investigação criminal, é crucial. O economista se preocupa com interpretabilidade porque é, de certa maneira, a finalidade principal do trabalho aplicado de métodos quantitativos. Estimar efeitos marginais, (semi)elasticidades e grandezas similares representa a esmagadora maioria das aplicações de econometria, salvo raros estudos como Edison e Carcel (2020) que usam técnicas não-supervisionadas vindas de Linguística Computacional.

Sobre a organização desta monografia. No capítulo 2 apresento a construção de um modelo de floresta aleatória partindo de primeiros princípios. No capítulo 3 discuto a teoria clássica de regressão linear, suas hipóteses, virtudes e limitações, amparado por simulações de Monte Carlo. Bem como a identificação de efeitos marginais em modelos lineares e a problemática envolvida em computa-los em modelos de floresta aleatória. No capítulo 4 apresento um estudo de caso ilustrando estimação, validação e interpretação de modelos de floresta aleatória, bem como uma ilustração da técnica. Por fim, no capítulo 5, ressalto limitações e agendas de pesquisa a partir daqui.

2 FLORESTAS ALEATÓRIAS

"Todos os modelos estão errados, mas alguns são úteis." - George Box

Neste capítulo desenvolvemos o modelo de floresta aleatória partindo de primeiros princípios. Primeiro veremos algumas definições e conceitos centrais ao Aprendizado de Máquina supervisionado, depois os conceitos de teoria dos grafos que sustentam Árvores de Decisão e discutir brevemente sua estimação. Por fim veremos quais propriedades indesejáveis desses modelos nos levam a construir florestas aleatórias.

2.1 Noções de Aprendizado de Máquina Supervisionado

Primeiro precisamos de algumas definições caras ao contexto de Aprendizado Supervisionado. Uma **observação** é um par de realizações de duas variáveis aleatórias. O par consiste em $x \in \mathbb{R}^k$, contendo k **variáveis explicativas** e $y \in \mathbb{R}$ que chamaremos de **variável resposta**. Com algumas convenções é possível representar dados não-numéricos. Variáveis binárias podem ser representadas por pares 0 e 1, então usar 'apenas' números reais não é uma grande limitação. O conjunto de vetores possíveis de serem medidos, o produto cartesiano de todos os espaços amostrais das variáveis aleatórias contidas no vetor com variáveis explicativas, é o **Espaço de Mensuração** $\mathcal{X}$ e $\mathcal{Y}$ o **Espaço de Resposta**, equivalente ao espaço amostral da variável resposta. Por exemplo: se uma dimensão deste espaço representa idade, entendemos que seu suporte está nos inteiros positivos entre 0 e, digamos, 120. Se uma variável é uma categoria binária, então seu suporte está em 0 e 1.

Munido de dados relacionando mensurações de variáveis e explicativas e com uma resposta a ser estudada, um modelador está interessado em um algoritmo que relacione medidas dentro de $\mathcal{X}$ com previsões para valores plausíveis de $\mathcal{Y}$. Essa representação matemática representando regras de previsão é o objeto fundamental do Aprendizado de Máquinas.

Definição 1 (Modelos) *Um **modelo** é uma função $f : \mathcal{X} \to \mathcal{Y}$. Se $\mathcal{Y}$ é enumerável dizemos que é um modelo de **Classificação**, se for um subconjunto convexo da reta real dizemos que é de **Regressão** e, em particular no caso em que $\mathcal{Y} = [0, 1]$, dizemos ser de **Risco**. A derivada $\frac{\partial f}{\partial x}$, $x \in \mathcal{X}$, se existir, é dita o **efeito marginal** da variável em relação a que se diferencia f.*

Classificar um paciente como sendo ou não portador de uma doença é um problema de classificação binária. Classificar o tipo de câncer a partir de medidas de um tumor é um problema de classificação multiclasse. Estimar o preço de um imóvel com base em suas características é um problema de regressão. Estimar a *probabilidade de default* de um empréstimo é um problema de risco, embora estimar se o empréstimo irá ou não entrar em default seja, novamente, um problema de classificação binária.

Ao definir um modelo não fizemos grandes hipóteses sobre f, apenas que se for diferenciável, sua derivada é uma grandeza de interesse. Temos infinitos modelos para qualquer problema de modelagem não-trivial, todos candidatos válidos. Dentro dessa enorme variedade de modelos alguns são florestas aleatórias, outros são de regressões lineares. O que se faz ao estimar um modelo é escolher, na classe de modelos que o procedimento de estimação comporta, o mais apropriado aos dados. Ao estimar uma árvore de decisão jamais obteremos um modelo de regressão linear, e vice-versa. Exatamente o que configura um modelo apropriado aos dados é uma questão que será discutida mais à frente.

Algumas propriedades interessantes de um modelo, como, por exemplo, são trivialmente obtidas em modelos de regressão linear. Basta diferenciar um polinômio. Essa propriedade não vale em funções definidas em trechos, como veremos que uma Floresta Aleatória é. Efeitos parciais são centrais para várias aplicações de econometria e demandam uma maneira de aproxima-los em outras classes, potencialmente mais interessantes que os lineares, de modelos.

2.2 Um pouco de Teoria dos Grafos

Vamos estabelecer alguns fatos básicos de Teoria dos Grafos, caracterizar árvores nesse contexto e oferecer um resultado com duas condições razoavelmente fracas e suficientes para estabelecer que um grafo é uma árvore. A apresentação seguirá Chartrand e Zhang (2013). Para um tratamento alternativo e mais amplo de Teoria dos Grafos o leitor pode verificar Bollobás (2013).

Definição 2 (Grafos) *Um **grafo** é um par $G = (V, E)$, onde V é um conjunto de elementos que chamamos **vértices** e os de E, **arestas**. Se uma aresta conecta dois vértices, dizemos ser aresta incidente aos vértices. Notamos o conjunto de vértices incidentes a uma aresta h pela função $\phi(h)$. O número de arestas que se liga a um vértice v é chamado de seu **grau**.*

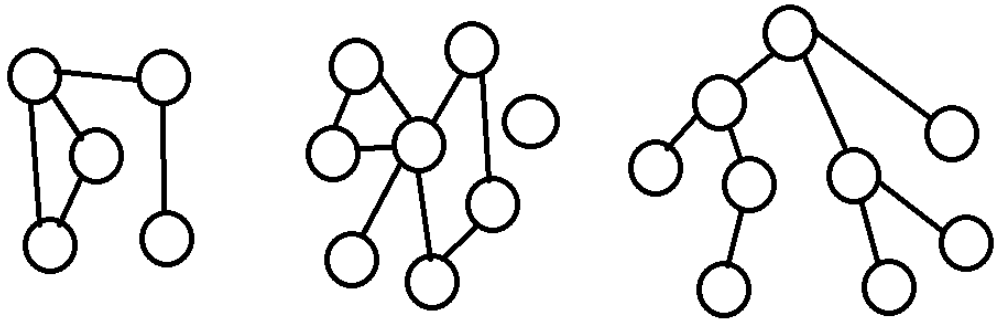
Definição 3 (Rotas em Grafos) *Um **passeio** é qualquer sequência de arestas $(h_1, h_2, ..., h_{n-1})$ para os quais há uma sequência de vértices $(v_1, v_2, ..., v_n)$ de forma que $\phi(h_i) = \{v_i, v_{i+1}\}$. Uma **trilha** é um passeio em que toda aresta é distinta. Um **caminho** é uma trilha em que todo vértice é distinto. Um **ciclo** é qualquer trilha que comece e termine no mesmo vértice. Um grafo que não admite ciclos é dito **acíclico**.*

Definição 4 (Conexidade) *Um grafo G é dito **conexo** se para qualquer par de vértices $x, y \in V \subset G$ há pelo menos um caminho cujo vértice inicial é x e o terminal é y. Um subconjunto de vértices de um grafo desconexo em que vale esta propriedade é dito um **componente**.*

Definição 5 (Árvores) *Um grafo G é dito uma **árvore** se para quaisquer dois vértices de G existe um caminho único os ligando. Podemos escolher um vértice arbitrário e defini-lo como a **raiz** da árvore. Os vértices que não são a raiz e têm grau unitário são ditos **folhas**. Os com grau*

maior que 1 *que não são a raiz são chamados* ***nodos****. Um conjunto disjunto de árvores é dito uma* ***floresta****.*

Figura 1 – Da esquerda para a direita: grafo conexo, um desconexo e uma árvore.



Fonte – Elaboração Própria

A Figura 1 ilustra algumas das definições anteriores. O resultado a seguir é interessante para nossa aplicação porque ilustra duas propriedades úteis de árvores.

Teorema 1 *G é uma árvore se, e somente se, é conexo e acíclico.*

Demonstração 1 *Seja G uma árvore. G é trivialmente conexo, pois por definição existe um caminho entre qualquer par de vértices. Suponha por absurdo que G admita um ciclo. Então existe uma trilha começando e terminando em um vértice v de G. Escolha um vértice qualquer u desse ciclo. Então existe um caminho $v \to u$ e uma trilha (possivelmente um caminho) $u \to v$. Dois casos ocorrem:*

- *Se $u \to v$ é uma trilha, é também a união de dois caminhos $u \to w$ e $w \to v$. Note que nesse caso podemos truncar o caminho $v \to u$ em w e estabelecemos dois caminhos distintos entre v e w. Em contradição com G ser uma árvore.*
- *Se $u \to v$ é um caminho então a contradição é imediata, pois existiriam dois caminhos entre u e v.*

Agora a volta. Tome G conexo e acíclico, escolha dois vértices v e u. Suponha por absurdo que exista mais de um caminho entre v e u. Então necessariamente existe ciclo no grafo G, basta 'ir' por um caminho e 'voltar' por outro. G é uma árvore. ■

2.3 Construindo Modelos com Árvores

Vamos agora cobrir duas situações hipotéticas que ilustram como essas definições aparecem em um contexto mais aplicado.

2.3.1 Classificação Binária

A unidade de email marketing de uma grande empresa de e-commerce quer segmentar clientes entre entusiastas de tecnologia, que engajarão felizmente com campanhas de aparelhos novos, e usuários relutantes de tecnologia, que não precisam receber esse contato nas suas caixas de email. Uma equipe selecionou algumas centenas de clientes aleatoriamente e manualmente os classificou usando entrevistas e análise de histórico de compras, um processo caro, demorado e de profundidade. Cabe agora a um analista tentar reproduzir os esforços manuais e não-escaláveis da equipe em um modelo preditivo que pode ser aplicado na base verdadeira de clientes usando os resultados do estudo.

O analista consultou o banco de dados da empresa e montou uma amostra contendo os seguintes dados: idade do cliente, percentual das compras em eletrônicos, valor médio da compra. Estamos falando de um espaço de mensuração $\mathcal{X} \subset \mathbb{R}^3$. Como a variável resposta é binária, $\mathcal{Y} = \{0, 1\}$.

Um procedimento possível seria primeiro estimar por mínimos quadrados generalizados um modelo para probabilidade uma observação pertencer a uma classe $f : \mathbb{R}^3 \to [0,\, 1]$ e depois alguma maneira de traduzir uma probabilidade presumida pelo modelo a uma classe $g : [0,\, 1] \to \{0, 1\}$. O modelo final, neste caso, seria a composição $g \circ f : \mathbb{R}^3 \to \{0, 1\}$.

Modelos têm hipóteses. Nesse caso uma delas é que as variáveis explicativas são independentes, no sentido de que sua distribuição conjunta é apenas o produto de suas distribuições marginais. A intuição do analista diz haver interações entre renda e idade, por exemplo. Um público mais velho que faz grandes compras em eletrônicos provavelmente é tão entusiasta quanto estudantes universitários que compram quase exclusivamente eletrônicos. O analista então pode seguir por outro caminho e enumerar uma série de perguntas sobre cada observação antes de emitir um julgamento:

- O cliente tem menos de 30 anos?
- Mais de um quarto das compras desse cliente foram em eletrônicos?
- A compra média desse cliente é maior que 250 reais?

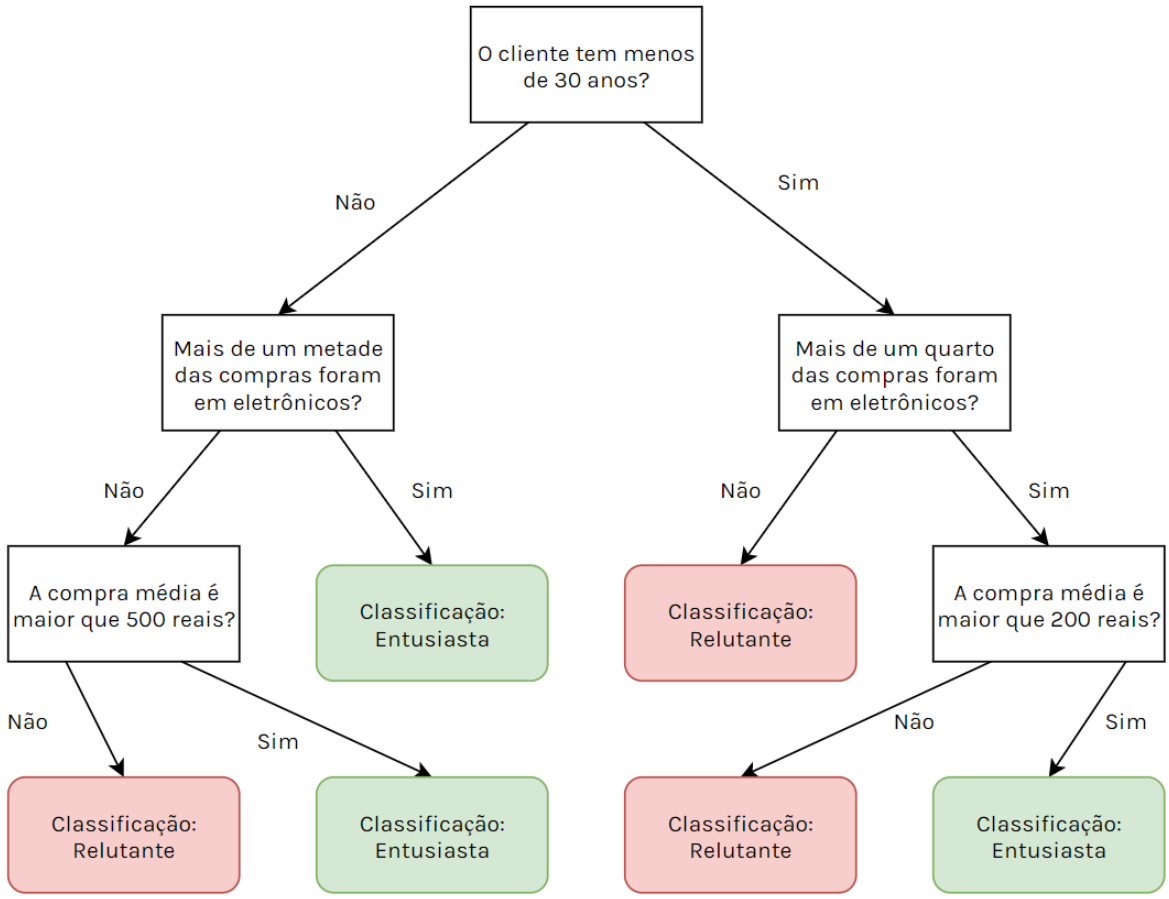
Cada pergunta aqui pode ser lida como uma função. 'O cliente tem menos de k anos?' é, computacionalmente, $f : \mathbb{R}^2_+ \to \{0, 1\}$. Algumas perguntas são mais informativas que outras. Afinal, um usuário teve mais de 90% das compras em eletrônicos é quase certamente um entusiasta de tecnologia, enquanto saber que um usuário tem mais de 20 anos provavelmente não é tão informativo.

Ao compor perguntas sucessivamente chegamos a uma espécie de fluxograma de decisão. Uma boa previsão de perfil para clientes, por exemplo, com menos de 30 anos que compram eletrônicos em 60% das compras passadas e gastam 40% a mais que a média por compra

provavelmente é que são usuários ávidos de tecnologia. Ao passo que um usuário de 72 anos que teve 10% das compras em eletrônicos e faz compras 40% menores que a média provavelmente não. Note, no entanto, que um usuário idoso que compra apenas eletrônicos na plataforma muito provavelmente é entusiasta. A composição das repostas é importante.

Supondo que toda pergunta seja sim ou não e que toda observação tenha uma resposta para toda pergunta, acabamos com uma estrutura recursiva. Nenhuma resposta individual altera as outras. Podemos elaborar com isso um fluxograma que tem a estrutura de um grafo árvore, representado graficamente na Figura 2.

Figura 2 – Um exemplo de árvore de decisão.



Fonte – Elaboração Própria

Recapitulando: a composição de perguntas e suas combinações distintas de respostas levam a uma partição do espaço de mensuração. Cada partição é associada a uma regra de previsão/ classificação. Podemos representar essa lógica com uma árvore. O modelo resultante é a função definida por trechos que atribui a cada região do espaço de mensuração uma regra de previsão. Formalmente:

Definição 6 (Modelos de Árvore) *Seja $\mathcal{X}, \mathcal{Y}$ um par de espaços de mensuração e resposta. Uma **Árvore de Decisão** é uma função definida por trechos $\mathcal{A} : \mathcal{X} \to \mathcal{Y}$, onde $\mathcal{Y}$ é enumerável.*

No caso em que $\mathcal{Y}$ ao, ao invés de enumerável, um subconjunto convexo da reta chamamos ***Árvore de Regressão****. Se $\mathcal{Y} = [0, 1]$ é comum se referir a uma* ***Árvore de Risco****.*

2.3.2 Regressão

Uma empresa de tecnologia no setor imobiliário quer trabalhar em um modelo de precificação de aluguel. A ideia é que donos de imóveis recebam um valor sugerido compatível com o mercado e embutir isso no serviço.

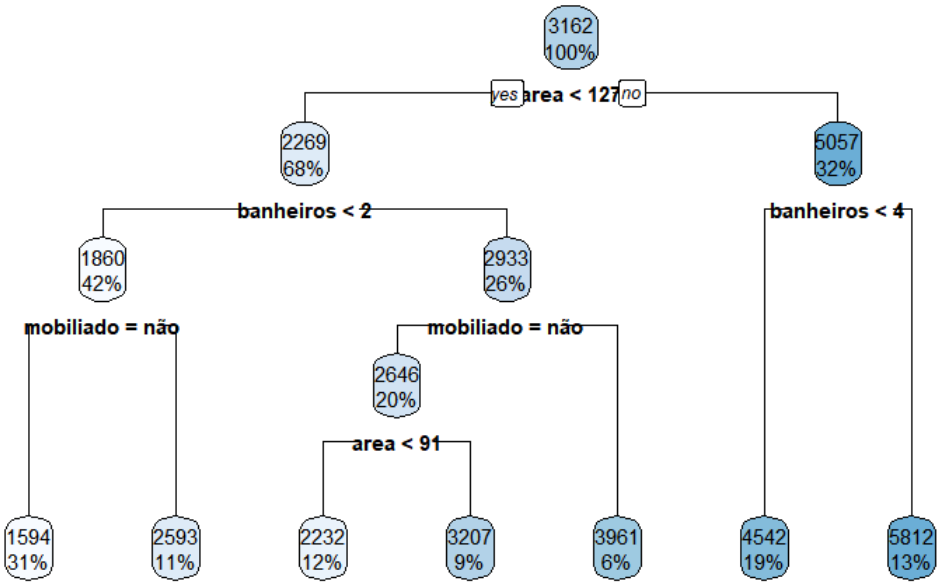
Um analista coletou dados de aluguel e algumas informações básicas do apartamento. Área, número de quartos, de banheiros, se aceita animais, se é mobiliado e em qual cidade está localizado. A Tabela 1 contém médias para as variáveis mais relevantes, por cidade.

Tabela 1 – Médias de variáveis por cidade.

Cidade	Área (m^2)	Quartos	Banheiros	Andar	Aluguel
Belo Horizonte	136.7	2.9	2.2	3.5	2765.9
Porto Alegre	93.3	2.1	1.7	3.9	2069.9
Rio de Janeiro	93.1	2.2	1.7	5.2	2774.7
São Paulo	124.7	2.4	2.2	5.6	3600.3

Fonte – Cálculos do autor com dados extraídos do site https://quintoandar.com.br

Figura 3 – Um exemplo de árvore de regressão.



Fonte – Elaboração Própria

A Figura 3 mostra a árvore de regressão estimada. As porcentagens se referem à fração da amostra original que passa pelos testes até ali, o número é a média da variável resposta no

grupo. Com as regras de previsão que estimamos, um apartamento com mais de 127 m^2 e mais de 4 banheiros tem aluguel presumido de 5812 reais. Já o aluguel de um com área menor, menos de 2 banheiros e sem mobília é presumido em 1594 reais.

2.4 Noções de Estimação de uma Árvore de Decisão

Nessa seção veremos como traduzir uma amostra A em uma árvore de decisão $\mathcal{A}$ Uma abordagem candidata é o procedimento original apresentado na primeira edição de Breiman et al. (1984). Imagine que temos uma massa de dados, uma nuvem de pontos em algum espaço de mensuração. Vamos elaborar uma enorme lista de sequência de perguntas, avaliar qual sequência é mais informativa e usar a sua árvore como modelo. Podemos avaliar o quão informativa é uma sequência de perguntas com partição recursiva:

Definição 7 *Um algoritmo de divisão de conjuntos que opere dividindo em subconjuntos os resultantes no passo anterior é dito uma* ***Partição Recursiva****. Toda partição recursiva pode ser representada com uma árvore. O número perguntas a ser feito, a métrica de qualidade de cada pergunta e, opcionalmente outros parâmetros como um critério para um valor mínimo da métrica de qualidade para aceitar uma pergunta, são ditos* ***hiperparâmetros****.*

Cada sequência de perguntas pode ser entendida como uma sequência de testes τ_i, a partição induzida pelas perguntas sendo representada por uma árvore $\mathcal{A}$ onde todo vértice que não seja uma folha tem um τ_i associado. Estimar a árvore é encontrar quais são as perguntas mais apropriadas, dado uma amostra.

Começamos com a amostra completa e enunciamos uma série de perguntas sobre ela — um procedimento computacionalmente intensivo por isso a estimação é quase sempre estocástica. Um subconjunto aleatório das perguntas possíveis é testado. Escolhemos a pergunta que melhor performa e partimos para os nodos "filhos". Repetimos o processo em cada um até algum gatilho ser ativado, então usamos o grupo para definir a regra de classificação/previsão. Em classificação podemos usar a classe mais comum no grupo, em regressão podemos usar a média ou mediana da variável resposta no grupo. Três gatilhos usuais:

- **Nenhum teste cumpre um valor mínimo para a métrica de qualidade**
  Cada pergunta respondida diminui a informação disponível, tornando o ganho de informação decrescente. Alguma hora, a cargo do modelador definir, uma pergunta adicional provavelmente custa mais em parcimônia e computação do que devolve em acurácia de previsão. É hora de gerar uma folha.
- **O número de perguntas feitas chegou ao máximo**
  Uma maneira de forçar parcimônia do modelo é limitar o número de perguntas. Algumas implementações diferenciam o número total de perguntas do número de perguntas sucessivas.

- **O número de observações do nodo não cumpre um mínimo**
  A amostra que chegou no nodo está abaixo de um mínimo. Por ser um dos mais simples e ter uma relação diretamente proporcional com o número de regras de previsão/classificação distintos é um dos mais comuns na literatura.

É esse o procedimento. Dividimos a amostra guiados por alguma métrica de sucesso e usamos a divisão final como um modelo estimado contendo regras de previsão/classificação informadas pelos dados. Resta entender o que queremos dizer, rigorosamente, com um teste ser *melhor* que outro.

2.4.1 Métricas de Informação

Voltando ao primeiro exemplo, a pergunta ”usuário tem mais de 90% das compras em eletrônicos?” provavelmente segrega muito bem as classes. É difícil conceber que um usuário assim não seja entusiasta de eletrônicos, ou pelo menos sustente alguém que é. Ela não é, no entanto, parcimoniosa. O tamanho do grupo de clientes que retorna positivo para essa pergunta dificilmente será relevante por isso essa pergunta provavelmente não será um bom insumo para uma regra de classificação. Pense também no grupo que **não** retorna positivo. Ele é provavelmente pouco segregado. Uma pergunta ”melhor”, espera-se, renderia dois grupos bem segregados e de tamanhos não muito díspares.

Começando pelo contexto de classificação, suponha J classes. Uma primeira métrica que respeita esse equilíbrio entre divisão e parcimônia é o **Ganho de Informação** (KULLBACK; LEIBLER, 1951). Podemos expressa-lo como a diferença entre a entropia do nodo ”pai” da divisão e a soma ponderada da entropia nos nodos filhos.

O vetor $P(n_i) = (p_1, p_2, ..., p_J)$ representa as probabilidades de que um elemento aleatoriamente escolhido do grupo que chegou no nodo n_i seja da i-ésima classe. A entropia do nodo n_i é $H(n_i) = -\sum_{p_j \in P(n_i)} p_j \log_2 p_j$. Defina $H(n_i \mid a), H(n_i \mid b)$ como a entropia dos potenciais nodos filhos de n_i, dado um teste τ que gere dois grupos a e b. Defina $P_a(\tau), P_b(\tau)$ como a proporção de elementos do nodo pai que vai para cada filho dado um teste τ.

Então o ganho de informação por entropia $\mathcal{I}_E(\tau)$ é dado por $\mathcal{I}_E(\tau) = H(n_i) - P_a(\tau)\, H(n_i \mid a) - P_b(\tau)\, H(n_i \mid b)$. Em cada divisão, usando essa métrica, escolhemos $\tau^* = \arg\max \mathcal{I}_E(\tau)$. Outra métrica de classificação é a **Impureza de Gini** (STROBL; MALLEY; TUTZ, 2009), definida como $\mathcal{I}_G(\tau) = \sum_{i=1}^{J} p_i(1 - p_i) = 1 - \sum_{i=1}^{J} p_i^2$.

2.5 Construindo uma Floresta Aleatória

Uma limitação do modelo construído até aqui é que ele é inclinado a *overfitting*, quando o modelo performa muito bem nos dados em que treinou e muito mal em dados novos. O problema é notoriamente acentuado em contexto de regressão. Afinal, um modelo de árvore parcimonioso

terá algo entre 3 e 20 regras de previsão distintas, ao passo que um modelo linear cobre uma região convexa.

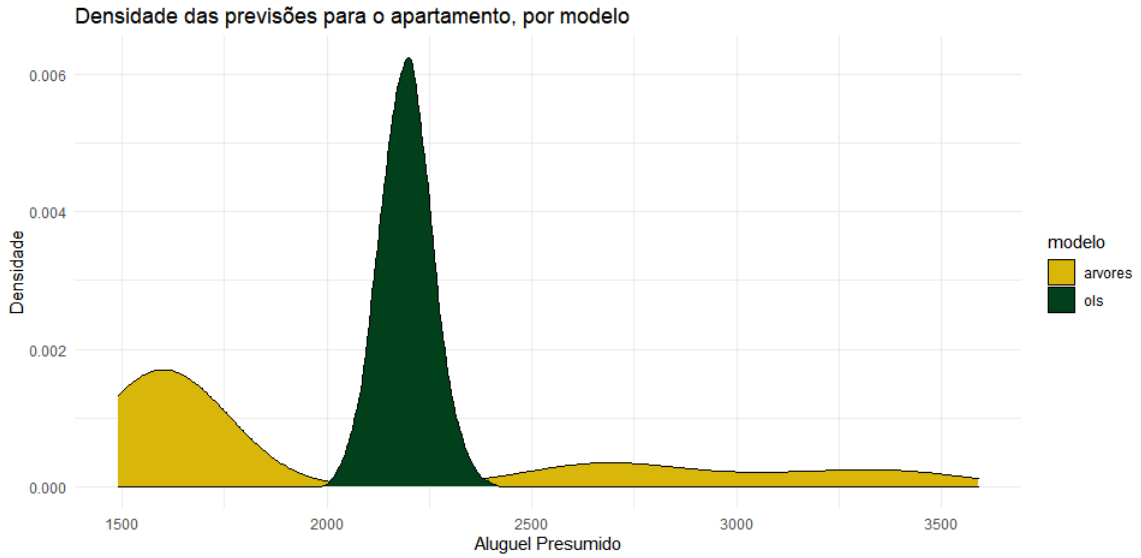
Isso deve a modelos de árvore serem modelos de alta variância: mudanças pequenas nos dados de treino podem gerar regras de previsão vastamente diferentes. Em compensação isso os torna modelos de baixo viés, capturam muito bem padrões nos dados apresentados. Ao contrário de modelos lineares, que tipicamente irão produzir previsões de variância mais baixa que modelos de árvore e com mais viés porque não conseguem incorporar relações latentes entre variáveis explicativas.

Essa troca entre viés e variância é uma espécie de restrição fundamental da modelagem e permeia Aprendizado de Máquina. O ***trade-off* viés-variância** é um conjunto de resultados para uma variedade de classes de modelos. Várias formulações desse resultado podem ser encontradas em Derumigny e Schmidt-Hieber (2020).

Podemos realizar uma simulação de Monte Carlo para por esse problema em perspectiva. No próximo capítulo exploraremos melhor modelos lineares e sua estimação por Mínimos Quadrados Ordinários (OLS). Agora, vamos ilustrar dois comportamentos desagradáveis dos modelos de árvore disponíveis até aqui, é para mitiga-los que vamos dar o próximo passo.

Voltando ao exemplo da árvore de regressão para preços de casas. Vamos agora selecionar aleatoriamente várias amostras de apartamentos e treinar uma árvore e um modelo linear (estimado por OLS) em cada. Como comparativo, vamos prever o aluguel de um apartamento de $82m^2$, 2 quartos, 1 banheiro, no Rio de Janeiro, não-mobiliado e aceite animais com cada um dos modelos e observar como as previsões se distribuem, como mostramos na Figura 4.

Figura 4 – A distribuição das previsões por classe de modelo.



Fonte – Elaboração Própria

O que acontece se ao invés de treinarmos **uma** árvore, treinar **várias** e usar a alguma

agregação das previsões individuais como previsão final? Em caso de classificação basta usar a moda, em caso de regressão basta usar a média. Sabemos pelo Teorema do Limite Central que a distribuição desse modelo seria normal, ao contrário do que acontece com a distribuição das previsões de uma árvore individual.

Teorema 2 (Lindenberg-Lévy) *Seja* $(\boldsymbol{X}_1, ..., \boldsymbol{X}_n)$ *uma amostra independente e identicamente distribuída com* $\mathbb{E}[\boldsymbol{X}_i] = \mu$ *e* $\mathbb{V}[\boldsymbol{X}_i] = \sigma^2 < \infty$*. Então:*

$$\sqrt{n}\left(\bar{\boldsymbol{X}}_i - \mu\right) \xrightarrow[n\to\infty]{d} N(0, \sigma^2) \tag{2.1}$$

Demonstração 2 *Ver Basu (1980).*

Agregar várias árvores não irá diminuir em absoluto a variância das previsões, temos modelos de baixo viés, mas as dará propriedades estatísticas mais amigáveis. Talvez melhor ainda, fará com que as previsões de um modelo de regressão passem a ocupar uma região convexa, assim como em modelos lineares. Podemos definir modelos de florestas aleatórias formalmente, como em Hastie, Tibshirani e Wainwright (2015) agora.

Definição 8 *Considere uma função* $\mu : \mathbb{R}^m \to \mathbb{R}$ *e uma floresta* $\mathcal{F}' = \{\mathcal{A}_1, ..., \mathcal{A}_m\}$*. Uma* ***Floresta Aleatória*** *é um modelo* $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$*,* $x \mapsto \mu(\mathcal{A}_1(x), ..., \mathcal{A}_m(x))$*.*

3 ECONOMETRIA CLÁSSICA E EFEITOS MARGINAIS

Introduzido o conceito de Floresta Aleatória, agora voltamos nossa atenção à Econometria Clássica e seu modelo central. Veremos uma maneira de estimar os parâmetros desse modelo via Mínimos Quadrados Ordinários, como modelos de Floresta Aleatória não têm a mesma interpretabilidade e como podemos contornar essa problemática avaliando o modelo na vizinhança de algumas observações. A apresentação seguirá Hayashi (2000).

3.1 Teoria Clássica de Regressão Linear

Retornando um pouco, e apenas esse pouco, ao capítulo anterior, usaremos de novo os conceitos de variáveis explicativas (que aqui chamaremos de **regressores**) e de uma variável resposta. Suponha que observamos n medidas. Então y_i é a i-ésima observação da variável resposta e o vetor $(x_{i1}, x_{i2}, ..., x_{ik})$ a i-ésima observação dos k regressores. Quando nos referimos a um modelo aqui, estamos nos referindo ao conjunto de restrições sobre a distribuição conjunta dos regressores e da resposta. A Teoria Clássica de Regressão Linear se apoia nas Hipóteses 1-4 a seguir.

Hipótese 1 (Linearidade) *Nossos modelos têm a seguinte forma funcional, onde β_j são os parâmetros a serem estimados e ϵ_i é o termo que chamamos de* ***resíduo****.*

$$y_i = \sum_{j=1}^{k} \beta_j x_{ij} + \epsilon_i; \tag{3.1}$$

Linearidade implica que o efeito de uma variação em um regressor particular na resposta não depende do seu nível, nem do de outros regressores. De fato:

$$\frac{\partial y_i}{\partial x_{ij}} = \beta_j \tag{3.2}$$

O modelador pode usar de intuição para construir variáveis novas que são funções não-lineares das variáveis mensuradas originalmente. A relação entre salário e experiência ou escolaridade, por exemplo, tem retornos decrescentes. Os primeiros cinco anos no mercado de trabalho contribuem muito mais para um aumento salarial do que os últimos cinco anos de carreira. Essa relação pode ser captada introduzindo um termo com o quadrado da experiência no modelo ou uma variável discreta valendo 1 nos primeiros anos de carreira e 0 depois, por exemplo.

Antes de prosseguir é importante apresentar a notação matricial dos modelos lineares. Uma maneira interessante de nos referir aos dados coletados de uma amostra - e como discutiremos estimação isso é importante - é associa-los à uma matriz. Notaremos uma **matriz de dados** como $\mathbf{X}$, preenchida com a i-ésima observação da j-ésima variável. Também teremos $\mathbf{y}$, o vetor em que a i-ésima entrada é a medida da variável resposta da i-ésima observação, e ϵ, o vetor com o resíduo. Finalmente, os parâmetros β_j estarão no vetor $\boldsymbol{\beta}$. A partir de agora usaremos n para nos referir ao tamanho da amostra, o número de linhas em $\mathbf{X}$ e $\mathbf{y}$. Reescrevendo a equação 3.1 em notação matricial:

$$\underset{n\times 1}{\mathbf{y}} = \underset{n\times k}{\mathbf{X}}\ \underset{k\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\epsilon}}. \tag{3.3}$$

Hipótese 2 (Exogeneidade Estrita) *A média condicional do erro é nula.*

$$\mathbb{E}[\epsilon_i \mid \mathbf{X}] = 0; \tag{3.4}$$

Essa hipótese não é restritiva se, entre as variáveis explicativas, houver uma com valor constante igual à média incondicional dos resíduos do modelo sem o valor constante. É assim de trás para frente que encontraremos o intercepto do modelo no processo de estimação, inclusive.

Hipótese 3 (Ausência de Multicolinearidade) *O posto da matriz de dados $\mathbf{X}_{n\times k}$ é k com probabilidade 1.*

Em termos práticos, supomos que as variáveis dadas para um modelo linear são linearmente independentes umas das outras. Se os valores de uma variável podem ser inteiramente determinados por combinações lineares de outras, qualquer informação que possa trazer já está contida nas outras. Também supomos que nosso modelo erra de maneira consistente:

Hipótese 4 (Homocedasticidade) *Seja $\mathbb{E}$ o operador de esperança:*

$$\mathbb{E}[\epsilon_i^2 \mid \mathbf{X}] = \sigma^2; \tag{3.5}$$

A variância dos resíduos independe do nível dos regressores.

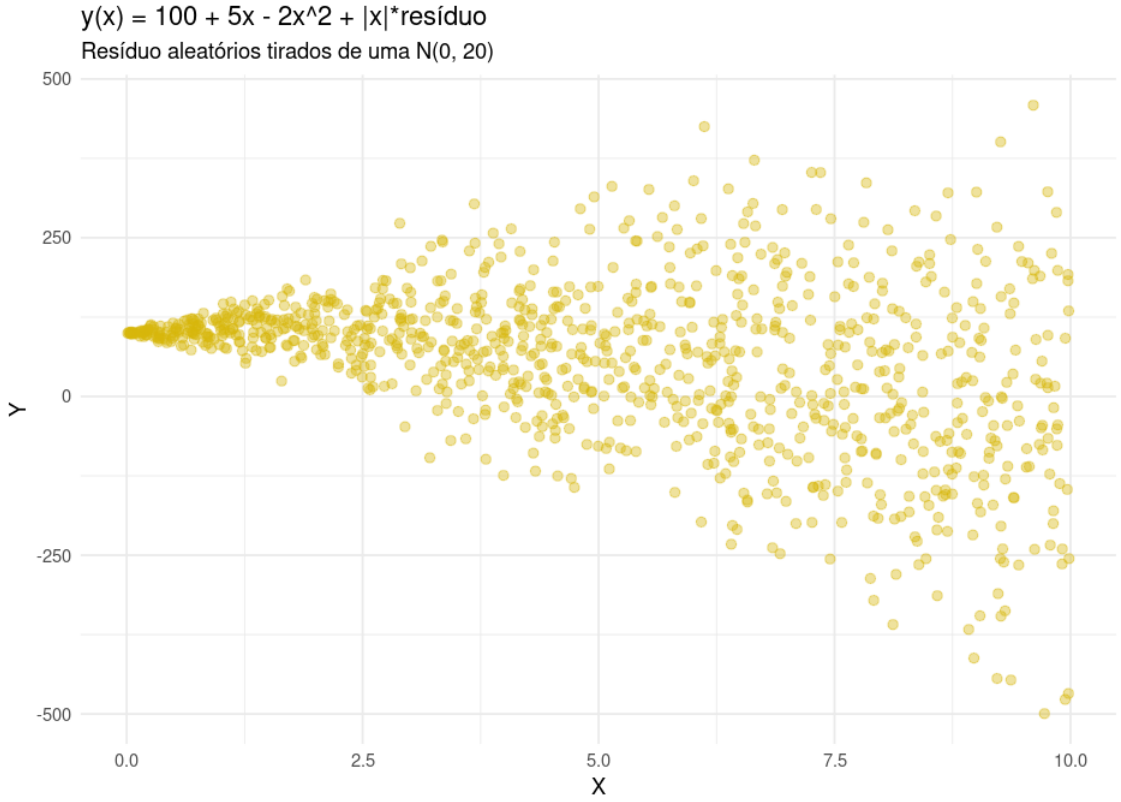
Exemplo 1 (Heterocedasticidade) *É simples violar a hipótese 4, basta tornar o componente não-observado uma função de alguma variável explicativa. Adicionando esse comportamento no processo simulado temos:*

Tabela 2 – Modelo com termo quadrático.

termo	estimativa	erro_padrao	estatistica_t
(Intercept)	97.64	6.97	14.01
x	24.72	3.15	7.86
x2	-2.37	0.30	-7.95

Fonte – Elaboração própria.

Figura 5 – Uma amostra simulada do processo.



Fonte – Elaboração própria.

Agora quando tentamos recuperar os parâmetros obtemos estimativas estatisticamente significantes, como informa a Tabela 2. Os parâmetros estimados, no entanto, estão à dezenas de desvios-padrão dos verdadeiros, como mostramos na Figura 6.

3.2 Efeitos Marginais

3.2.1 Em Modelos Lineares

A primeira hipótese, Linearidade, é a chave aqui. A resposta, supomos, varia linearmente nos regressores. Podemos introduzir algumas interações criando variáveis novas que são funções não-lineares das variáveis originais, adicionamos logaritmos, potências, indicadoras e outras tantas transformações para acomodar não-linearidades. Essa abordagem promissora, no entanto,

necessariamente diminui a precisão da estimativa de todos os parâmetros e seus graus de liberdade, impondo restrições.

Mesmo que essa perda de precisão seja tolerável, esbarramos em um problema embutido na construção desses modelos. Estimamos um escalar para descrever o efeito marginal de uma variável sobre a resposta e apenas isso. Estamos supondo por definição que os efeitos marginais são iguais em todo o suporte e para qualquer nível dos outros regressores. Essa hipótese, no entanto, não está presente em florestas aleatórias.

3.2.2 Em Florestas Aleatórias

Suponha que temos uma árvore treinada $\mathcal{A}$ e uma observação x. Com alguma licença poética nos referimos à previsão dessa árvore para essa obervação como $\mathcal{A}(x)$. De maior incômodo para o econometrista é que não é claro o que exatamente é a derivada dessa função. Isso é um problema porque boa parte da utilidade de um modelo (linear) estimado é ter um vetor de parâmetros intuitivamente interpretáveis, pois contém as derivadas parciais do modelo.

A derivada de $\mathcal{A}(\cdot)$, seja lá como for, não é tão informativa. Uma pertubação em uma observação x só altera o resultado da previsão de uma árvore se for grande o suficiente para deslocar x para outra regra de classificação/previsão. Teríamos uma função que é nula em boa parte de seu domínio, descontínua onde não for.

O problema é atenuado com uma floresta aleatória. Uma pertubação pequena em x pode alterar a previsão de uma fração das árvores da florestas. Com um número suficientemente grande de árvores uma variação arbitrariamente pequena em x leva à uma variação arbitrariamente pequena em $\mathcal{F}(x)$ e vale alguma forma de continuidade.

Esse caminho sugere uma estratégia promissora. Podemos perturbar uma observação de referência e as previsões de seus vizinhos. Isso gera uma curva relacionando valores de um regressor, dado um vetor níveis para os outros regressores, às previsões. Sua inclinação nos dá os efeitos marginais, que ao contrário do que acontece em modelos lineares, são sensíveis aos níveis dos regressores que não estamos perturbando.

3.3 Um Procedimento de Computação

Suponha um modelo $\mathcal{M} : \mathcal{X} \to \mathcal{Y}$. Como computamos seus efeitos marginais de maneira agnóstica? Gostaríamos de aplicar um mesmo procedimento e identificar os efeitos marginais sem depender da mecânica particular de uma classe de modelos. Por trás dos panos cada classe opera um maquinário completamente diferente: redes neurais são composições sucessivas de transformações lineares, modelos lineares são polinômios, máquinas de vetores de suporte calculam distâncias de vetores a um hiperplano estimado com base nos dados.

Se as mecânicas internas de $\mathcal{M}$ variam demais para termos um procedimento uniforme, podemos olhar onde não há variação entre classes de modelos, $\mathcal{X}$ e $\mathcal{Y}$. Observando os valores

que as predições do modelo dão para a resposta Y em uma curva parametrizada $t \in \mathcal{X}$ basta computar a sua derivada para achar os efeitos marginais na curva.

Como normalmente estamos interessados no efeito de tratamento individual de uma variável, talvez seu efeito conjunto com outra, o problema é limitado a avaliar o modelo ao longo de uma ou duas dimensões apenas. Avaliar o efeito marginal de mais variáveis implica apenas repetir o procedimento, não sofrer da maldição da dimensionalidade. Esse procedimento é computacionalmente simples, barato e agnóstico ao modelo.

4 APLICAÇÃO DO PROCEDIMENTO

Neste capítulo o procedimento será ilustrado em um modelo de regressão de preços de imóveis. Todos os dados e código de implementação estão disponíveis sob a licensa MIT no repositório https://github.com/pedrocava/monografiaRandomForest. Os dados foram adquiridas via *webscrapping* de um site nacional de aluguel de imóveis (https://www.quintoandar.com.br) para quatro capitais brasileiras no dia 20 de Março de 2020, realizado pelo autor. Iremos realizar análise exploratória para entender melhor os dados de alugueis, estimar uma série de modelos com variadas configurações de hiperparâmetros e computar algumas métricas de sucesso para explorar como eles afetam performance do modelo. Essas métricas de sucesso serão computadas com dados omitidos no processo de treinamento, num processo chamado Validação Cruzada, e então aplicaremos o procedimento descrito no capítulo anterior.

Antes, no entanto, veremos o procedimento ser aplicado em um contexto laboratorial. Simularemos dados aleatórios a partir de um modelo linear e recuperaremos o parâmetro estimado com florestas aleatórias e uma aplicação do procedimento.

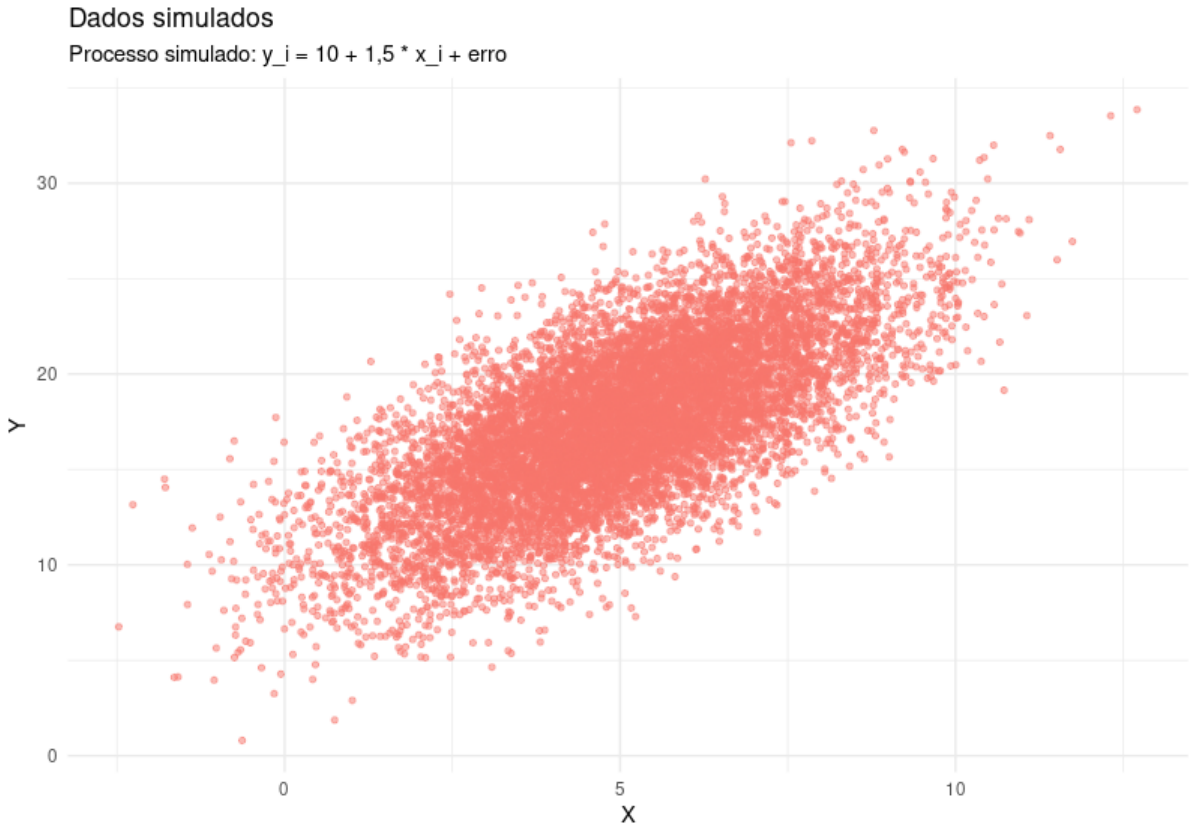
4.1 Verificação Laboratorial

Para uma primeira verificação da sanidade da técnica apresentada, o seguinte procedimento será feito. Primeiro dados aleatórios serão gerados a partir de uma normal. Esse será nosso regressor. Depois serão simulados outros dados de uma normal, que serão os erros. Depois, um cálculo determinístico irá gerar a nossa resposta simulada com um parâmetro conhecido.

Será treinada uma floresta aleatória nessa amostra falsificada e suas previsões serão computadas para variados valores do regressor imaginário. Iremos, então, estimar uma regressão linear do regressor simulado original nas predições da floresta aleatória. Caso a floresta de fato tenha conseguido recuperar o parâmetro, o β que escolhemos para o cálculo determinístico da resposta deverá ser o mesmo que encontraremos na segunda regressão.

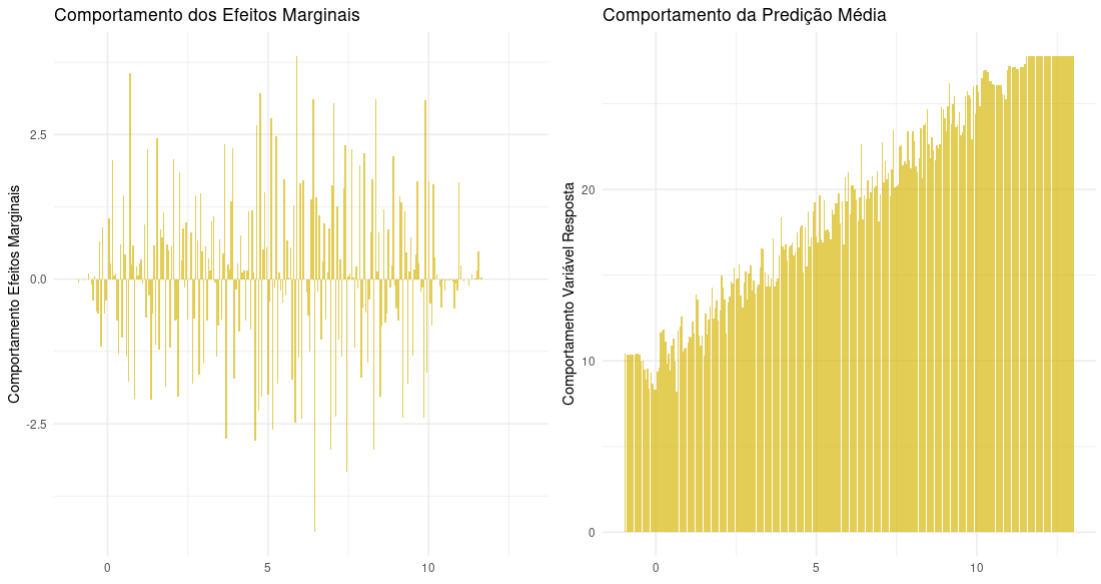
As Figuras 6, 7 e 8 descrevem, respectivamente: a distribuição conjunta dos dados simulados, o resultado da aplicação do procedimento de computação de efeitos marginais ao lado da curva de predição média e a distribuição conjunta dos valores preditos pelo modelo de floresta aleatória com os verdadeiros. Há uma relação aparentemente 1 para 1, o que é um sinal muito positivo já antes de verificar os resultados.

Figura 6 – Dados simulados.



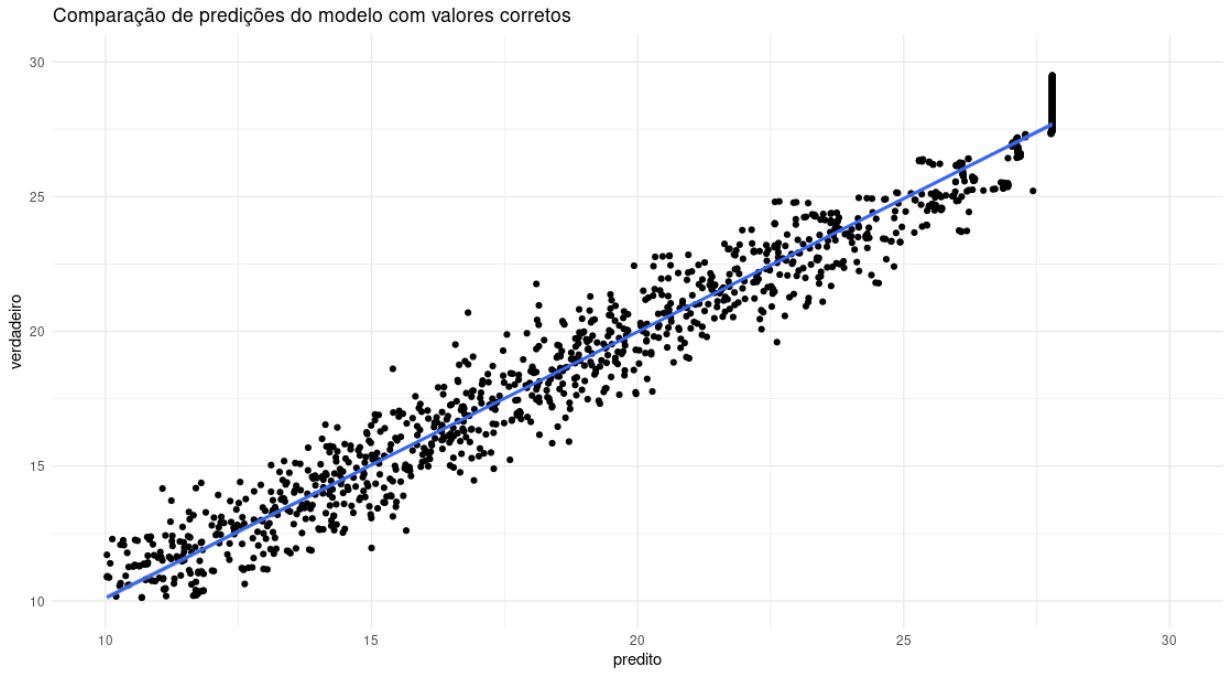
Fonte – Elaboração própria.

Figura 7 – Comportamento dos efeitos marginais, que lembra um ruído branco, e a curva de predições.



Fonte – Elaboração própria.

Figura 8 – Comparação das predições com os valores verdadeiros.



Fonte – Elaboração própria.

Na Tabela 3 estão os resultados para a segunda regressão do experimento. Ao gerar a resposta simulada, escolhemos um β de exatamente 1.5. Nosso procedimento recuperou esse parâmetro com um erro de cerca de $2,5\%$.

Tabela 3 – Resultados da verificação com dados simulados.

X	1.465^{***}
	(0.007)
Constante	10.243^{***}
	(0.050)
N	1,401
R^2	0.970
Adjusted R^2	0.970
F Statistic	$44{,}946.980^{***}$ (df = 1; 1399)
	*p<0.1; **p<0.05; ***p<0.01

Fonte – Elaboração própria.

4.2 Exploração dos Dados de Imóveis

Nesta seção algumas visualizações de dados úteis serão apresentadas, bem como uma descrição dos dados. Foram mensuradas as seguintes variáveis:

- **Cidade:** a cidade onde está o apartamento.
- **Área:** a área em metros quadrados
- **Quartos:** o número de quartos
- **Banheiros:** o número de banheiros
- **Vagas:** o número de vagas
- **Andar:** em qual andar do prédio está o apartamento (0 para térreo ou casas)
- **Mobiliado:** variável categórica indicando se o apartamento é mobiliado
- **Aluguel:** valor do aluguel em reais por mês

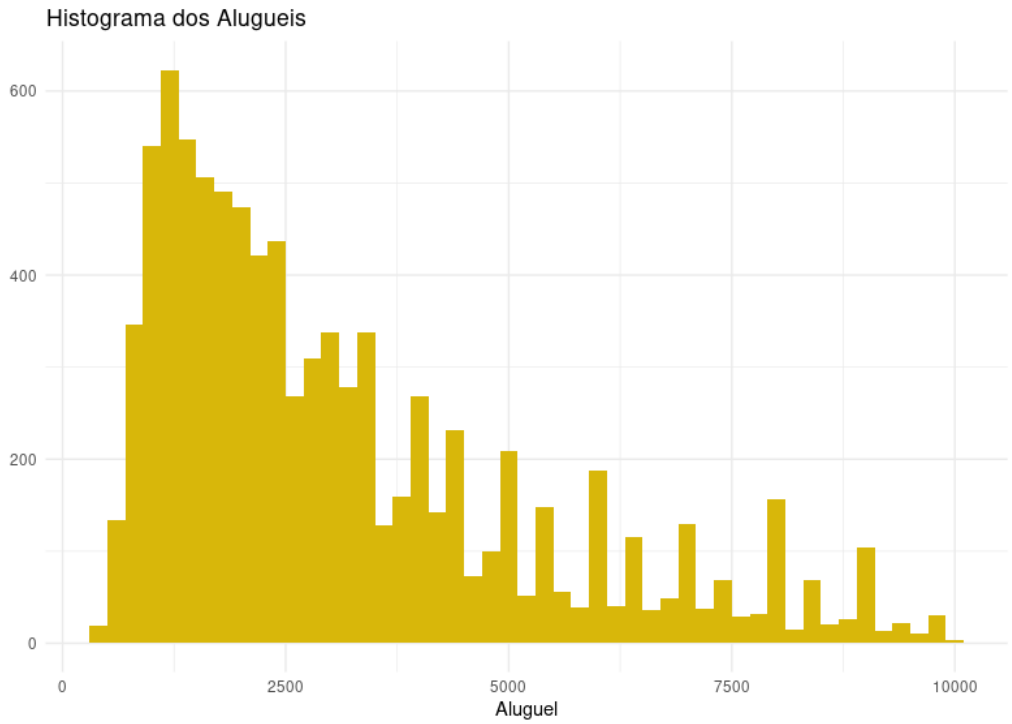
Na Tabela 4 algumas estatísticas descritivas importantes por cidade e alguns gráficos descrevendo as variáveis chaves.

Tabela 4 – Estatísticas descritivas por cidade.

Cidade	Área Média	Quartos / apt	Aluguel Médio	Custo Médio por M^2
Belo Horizonte	136.71	2.9	2765.90	22.81
Porto Alegre	93.27	2.0	2069.88	25.29
Rio de Janeiro	93.14	2.1	2774.70	34.18
São Paulo	124.69	2.4	3600.26	37.62

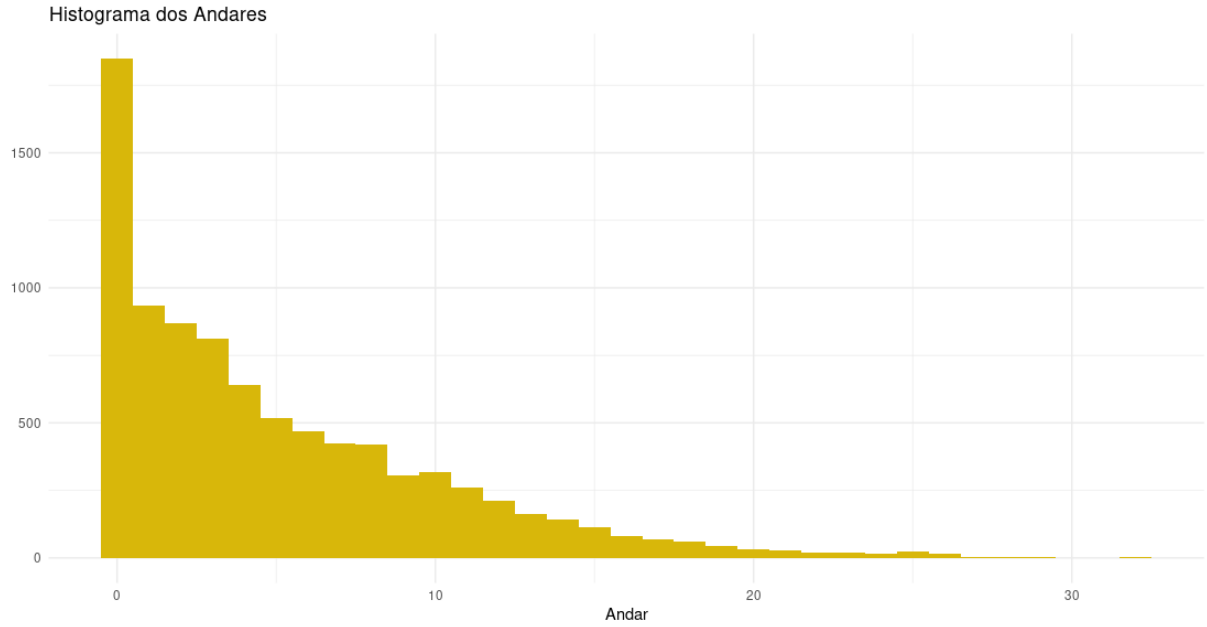
Fonte – Cálculos do autor com dados raspados do site https://quintoandar.com.br.

Figura 9 – Distribuição dos aluguéis.



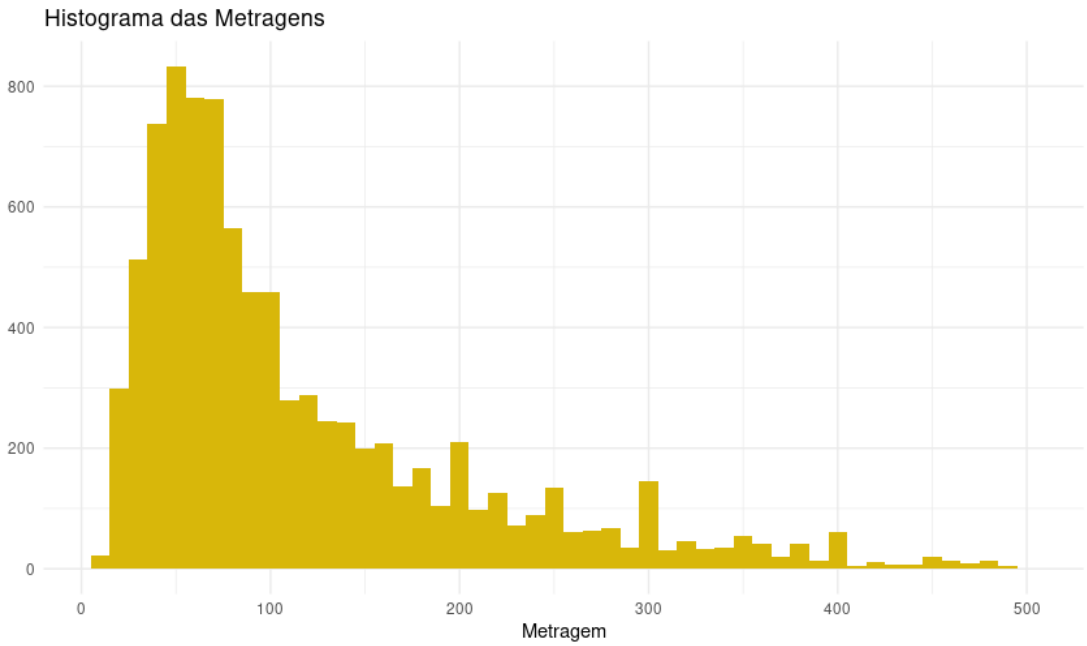
Fonte – Cálculos do autor com dados raspados do site https://quintoandar.com.br.

Figura 10 – Distribuição dos andares dos apartamentos.



Fonte – Cálculos do autor com dados raspados do site https://quintoandar.com.br.

Figura 11 – Distribuição da metragem dos apartamentos.



Fonte – Cálculos do autor com dados raspados do site https://quintoandar.com.br.

4.3 Otimização de hiperparâmetros

Não existe uma única maneira de estimar uma floresta aleatória. De fato, como machine learning é um campo que cresceu muito às margens da academia, em laboratórios da indústria, as convenções são informais e há pouco escrito em pedra. Optei por usar a implementação em Wright e Ziegler (2017), que usa como gatilho de geração de folhas uma amostra abaixo da mínima chegar no nodo e aleatoriza quantas variáveis explicativas são usadas em cada árvore. Uma alternativa de alta confiabilidade seria Liaw e Wiener (2002). Para a otimização de hiperparâmetros, validação cruzada e avaliação dos modelos foi usada a plataforma `tidymodels` (KUHN; WICKHAM, 2020), implementada em linguagem R (R Core Team, 2020).

Decidida a implementação que irá realizar as computações, é preciso fazer uma escolha sobre os hiperparâmetros do modelo, no caso três: amostra mínima para criação de folha, número de árvores e número de variáveis a serem aleatoriamente escolhidas para cada árvore. Do ponto de vista do expectador desinteressado o problema é apenas:

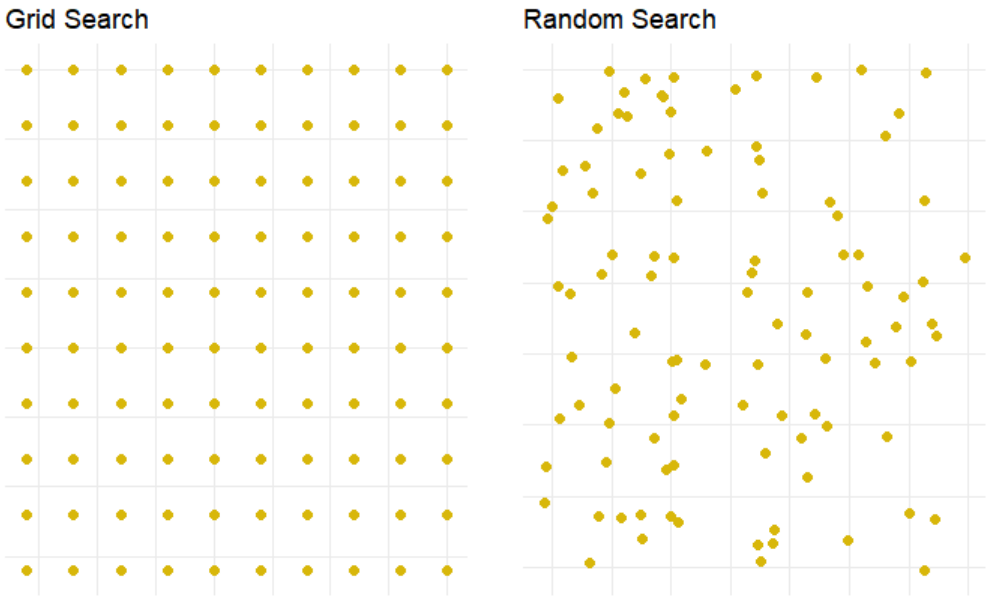
$$\mathbf{x}^* = \arg\max_{\mathbf{x}} \phi(\mathbf{x}) \tag{4.1}$$

E como seria fácil se soubéssemos exatamente o que é $\phi(\cdot)$, mas não sabemos. Uma primeira abordagem é tatear o suporte da função em busca da melhor combinação. Um método para isso é *random search*, gerar algumas combinações de hiperparâmetros aleatórias, estimar um modelo em cada e escolher o de melhor performance, que definiremos em detalhes brevemente.

O dual dessa abordagem é *grid search* em que ao invés de gerar combinações aleatórias se cobre o espaço de hiperparâmetros de vetores com distância regular, gerando uma grade.

Essas são ditas abordagens **caixa-preta livre de modelo** pois não fazem suposições sobre a forma funcional da função a ser maximizada ajustando os hiperparâmetros. Varremos o que acreditamos ser o seu domínio à força bruta. Existem alternativas, Shahriari et al. (2015) é um tratamento amplo da principal, otimização bayesiana. Otimização de Hiperparâmetros é um campo vasto. Um tratamento mais detalhado do estado da arte na área está disponível em Feurer e Hutter (2019).

Figura 12 – Duas buscas de 100 modelos cada.



Fonte – Elaboração própria.

4.4 Métricas de Qualidade

Na seção anterior optamos por escolher hiperparâmetros de forma a maximizar alguma função que entendemos representar a qualidade de um modelo. Precisamos agora definir como mensura-la. Métricas de performance de modelo são sempre arbitrárias então a boa prática recomenda usar uma cesta delas. As opções para regressão são inúmeras. Serão avaliadas seis:

- **Razão Performance-Interquatil (RPIQ)**
  Definida como o desvio-padrão das previsões dividido pela amplitude interquartil observada da resposta. Mede a capacidade do modelo de gerar previsões parcimoniosas. Quanto mais próximo de 1, melhor. Um tratamento mais detalhado está disponível em Botchkarev (2018).

- **Coeficiente de Concordância de Correlação (CCC)**
  Introduzido por Lawrence e Lin (1989), mede acurácia. Calculada a partir da diferença entre a identidade e a reta de regressão dos valores preditos nos observados. Quanto mais próximo de 1, melhor.

- **Erro Médio Absoluto (MAE)**
  A média das diferenças entre previsões e valores observados. Diretamente interpretável na unidade original da resposta. Indica acurácia. Menor é melhor, mas deve-se atentar à parcimônia.

- **Erro Médio Percentual (MPE)**
  A média das diferenças entre previsões e valores observados ponderada pela média da resposta. Lida em unidades relativas. Indica acurácia. Menor é melhor, mas deve-se atentar à parcimônia.

- **Coeficiente de Determinação (R2)**
  Calcula-se a soma dos quadrados dos desvios das previsões à média da resposta. Dividi-se esse valor pela soma dos quadrados dos desvios as observações originais à média da resposta. O resultado está entre 0 e 1 e pode ser interpretado como a fração da variância que o modelo consegue explicar. Maior é melhor, mas deve-se atentar à parcimônia, pois cresce monotonamente no número de variáveis explicativas.

  Uma variante dessa métrica, especialmente apropriada para comparar modelos com os mesmos hiperparâmetros (ou que não dependam deles, como regressão linear) e variáveis explicativas diferentes é o R2 ajustado. Seu cálculo é muito similar, mas introduz um mecanismo de punição à adição de variáveis, refletindo os benefícios da parcimônia na métrica.

- **Raiz do Erro Quadrático Médio (RMSE)**
  O Erro Quadrático Médio é a média dos quadrados dos desvios das previsões em relação aos valores observados. A raiz dá a métrica. Mede principalmente acurácia, embora seja muito sensível a outliers. Menor é melhor.

Em um contexto de classificação precisamos de outras métricas. A título de exemplo, algumas das mais amplamente utilizadas são:

- **Sensitividade/ Recall**
  Em classificação binária, a Sensitividade é a proporção dos casos positivos que são corretamente identificados.

- **Especificidade**
  Em classificação binária, a Especificidade é a proporção dos casos negativos que são corretamente identificados.

- **Acurácia**
  A fração de casos corretamente identificados.

- **Precisão/ Valor Preditivo Positivo**
  O número de observados positivos dividido pelo número de preditos positivos.

- **F1-Score**
  Média harmônia da Precisão e da Sensitividade

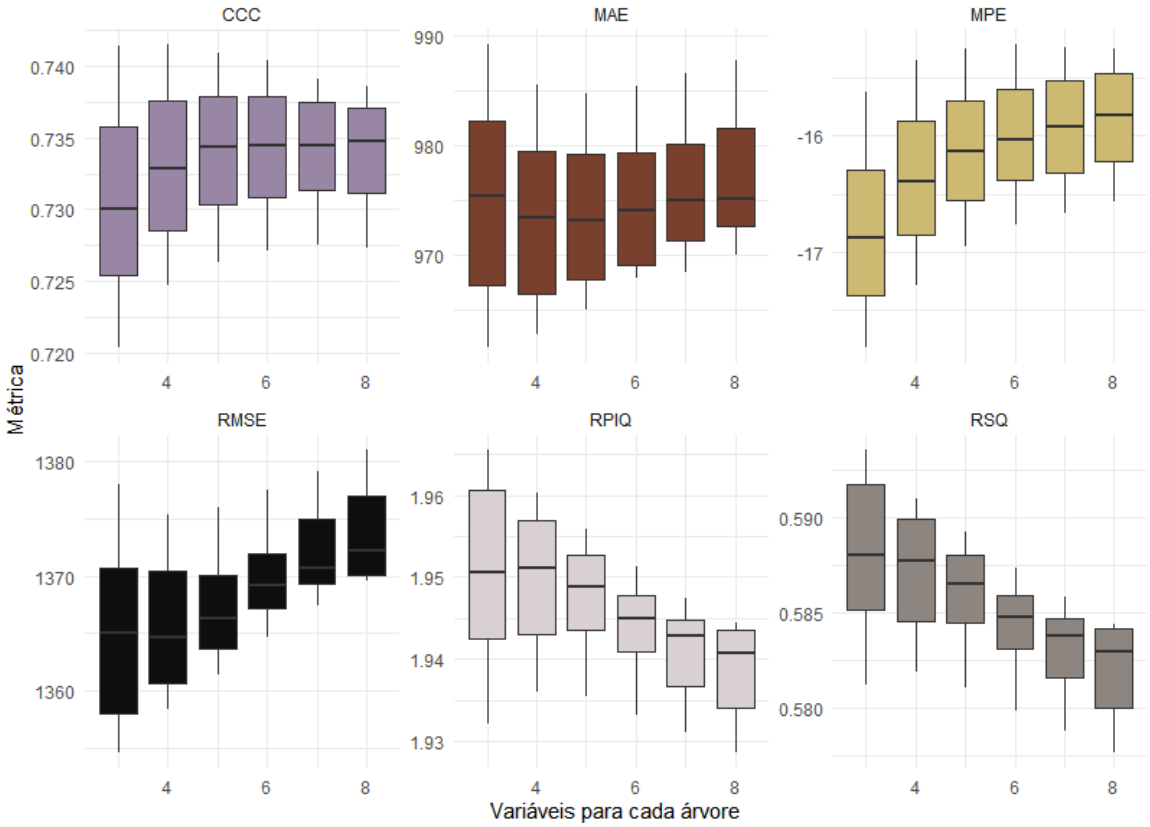
4.5 Validação Cruzada

Para dar uma chance melhor a cada combinação podemos testa-la em amostras diferentes. Fazemos isso com validação k-cruzada. Dividimos a amostra em k grupos aproximadamente iguais e para cada combinação de hiperparâmetros estimamos o modelo em $k-1$ combinações, excluindo um grupo de cada vez. Como medida de performance para cada combinação específica de hiperparâmetros usamos então a média das métricas de performance nas $k-1$ validações.

A literatura da área sugere que o número de árvores não é um bom hiperparâmetro para se validar (CLAESEN; MOOR, 2015). Os ganhos de performance são pequenos, o custo computacional, no entanto, variante. Note que o custo de estimar uma floresta cresce linearmente, um para um, com o número de árvores. Validar k combinações cada uma com $a * b$ árvores é a vezes mais caro que validar k combinações com b árvores.

Esse tempo de computação é melhor empregado procurando com uma grade fina melhores combinações de amostra única e número de variáveis por árvore. Árvores com menos variáveis tem menos variância, o que ajuda a diminuir a variância da floresta, e menor poder preditivo. Árvores com menor amostra mínima têm mais folhas, criando respostas mais finas, mas têm mais variância também.

Figura 13 – Resumo de métricas de performance variando quantas variáveis alimentar para cada árvore.



Fonte – Elaboração própria.

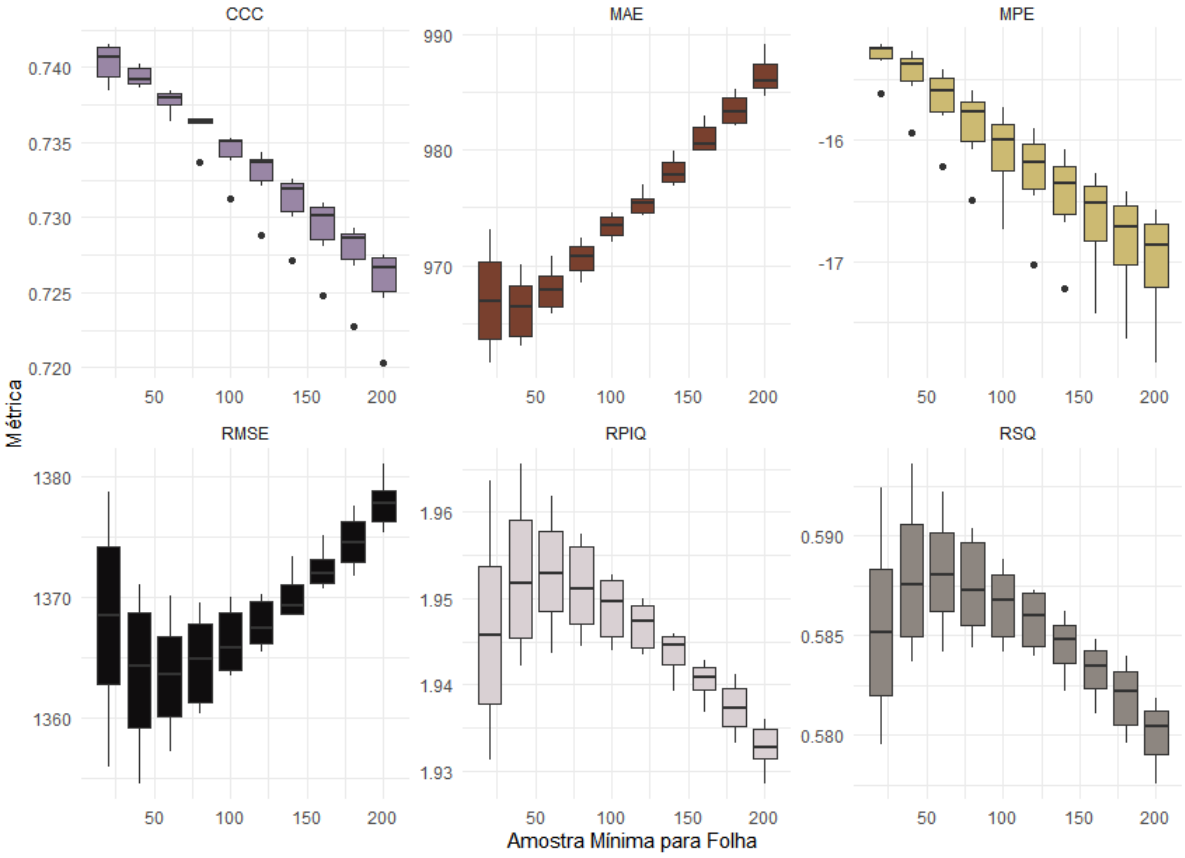
A varredura pelo hiperparâmetro do número de variáveis sugere que ele não é muito determinante para performance. Não há movimentos claros entre as métricas de melhora ou piora, como mostramos na Figura 13. Poucas variáveis por árvore incluem tanto os melhores quanto os piores modelos, como mostramos na Tabela 4. Isso sugere que a amostra mínima para gerar folha é um hiperparâmetro mais relevante. De fato é:

Tabela 5 – Melhor modelo de acordo com cada métrica.

Variáveis por Árvore	Amostra Mínima para Folha	Métrica
3	20	RPIQ
4	20	CCC
3	20	MAE
3	200	MPE
3	20	RSQ
3	20	RMSE

Fonte – Elaboração própria.

Figura 14 – Resumo de métricas de performance variando a amostra mínima para criar uma folha.



Fonte – Elaboração própria.

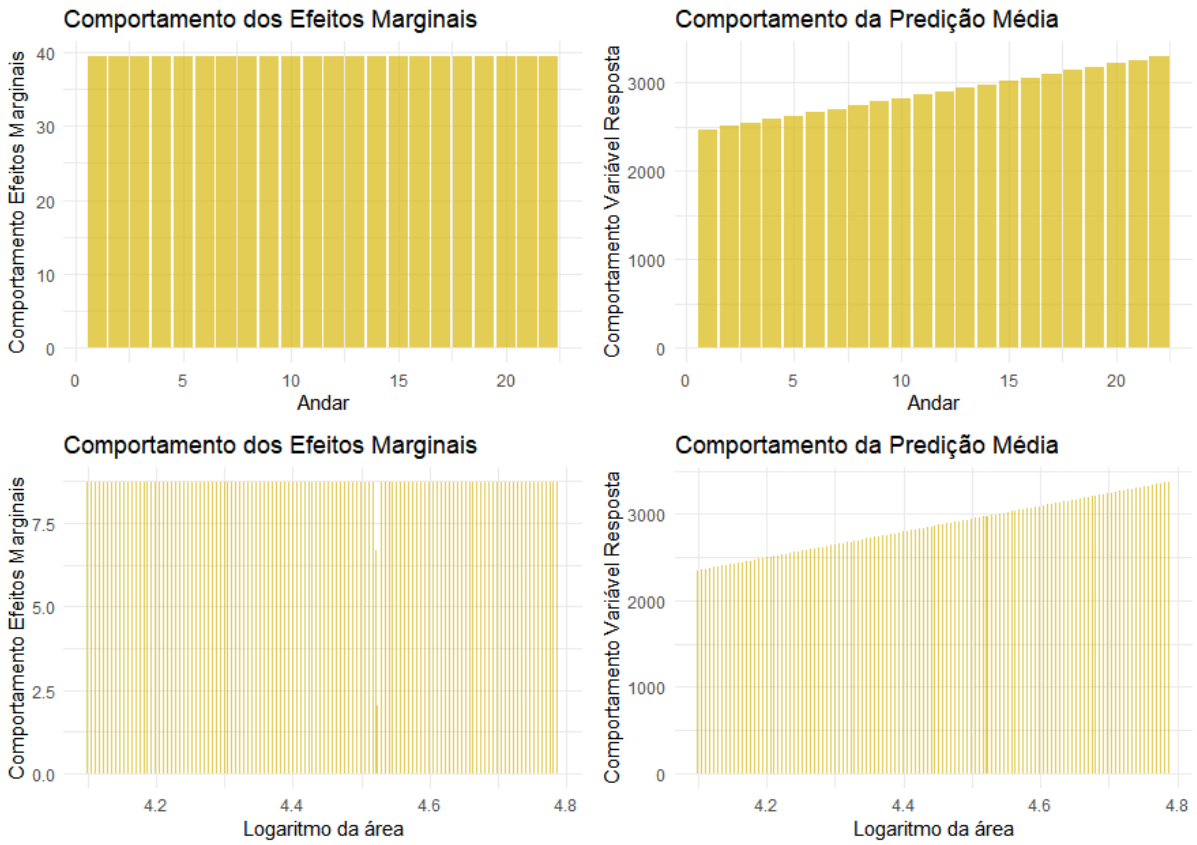
4.6 Computação de Efeitos Marginais

Primeiro, vamos definir uma referência. Nosso apartamento imaginário tem 92 m^2, 3 quartos, 2 banheiros, uma vaga na garagem, fica no oitavo andar, não é mobiliado, fica no Rio de Janeiro e o dono aceita animais. O que seria dele se fosse maior, ou em um andar mais alto?

Ao aplicar o procedimento com um modelo linear, o que temos? Justamente o que

modelos lineares deveriam devolver na ausência de termos com interações e funções não-lineares dos regressores originais, efeitos marginais constantes, mostrado na Figura 15.

Figura 15 – A aplicação do procedimento em modelos lineares ilustra a primeira hipótese dos modelos.

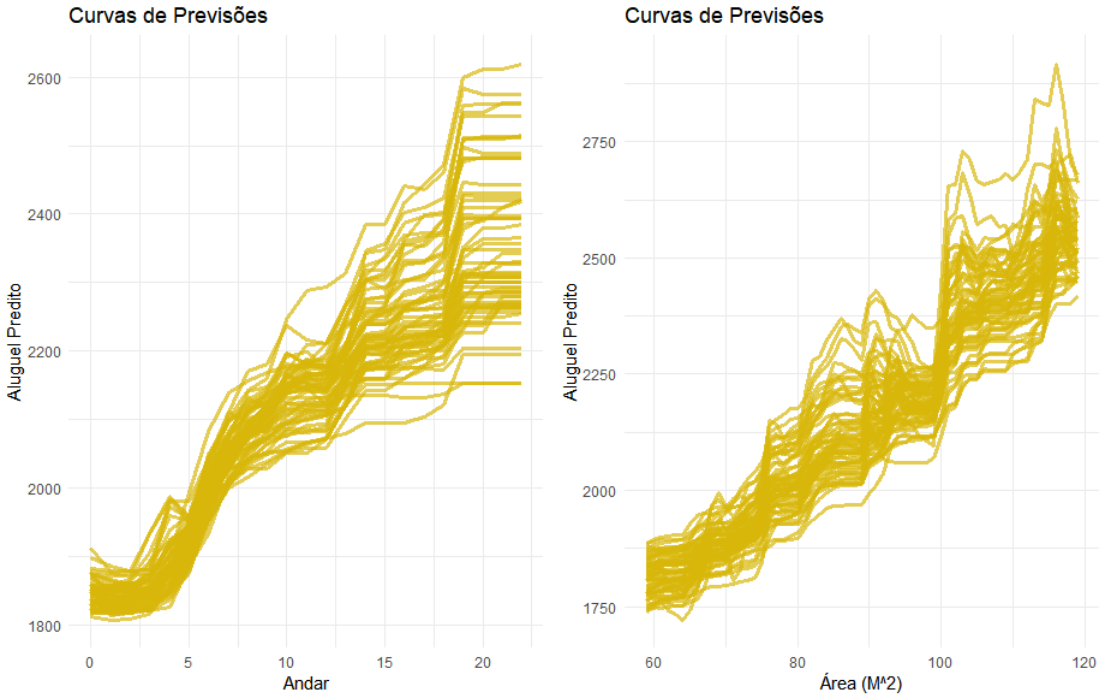


Fonte – Elaboração própria.

Já em modelos de floresta aleatória a não-linearidade do fenômeno fica evidente. Sair do térreo para o primeiro andar *barateia* um apartamento, presumivelmente porque o benefício de uma vista é quase inexistente embora o custo de subir escada/elevador apareça. Sair do décimo segundo para o décimo terceiro tem um efeito nulo e por vezes negativo. Possivelmente por conta da superstição com o número.

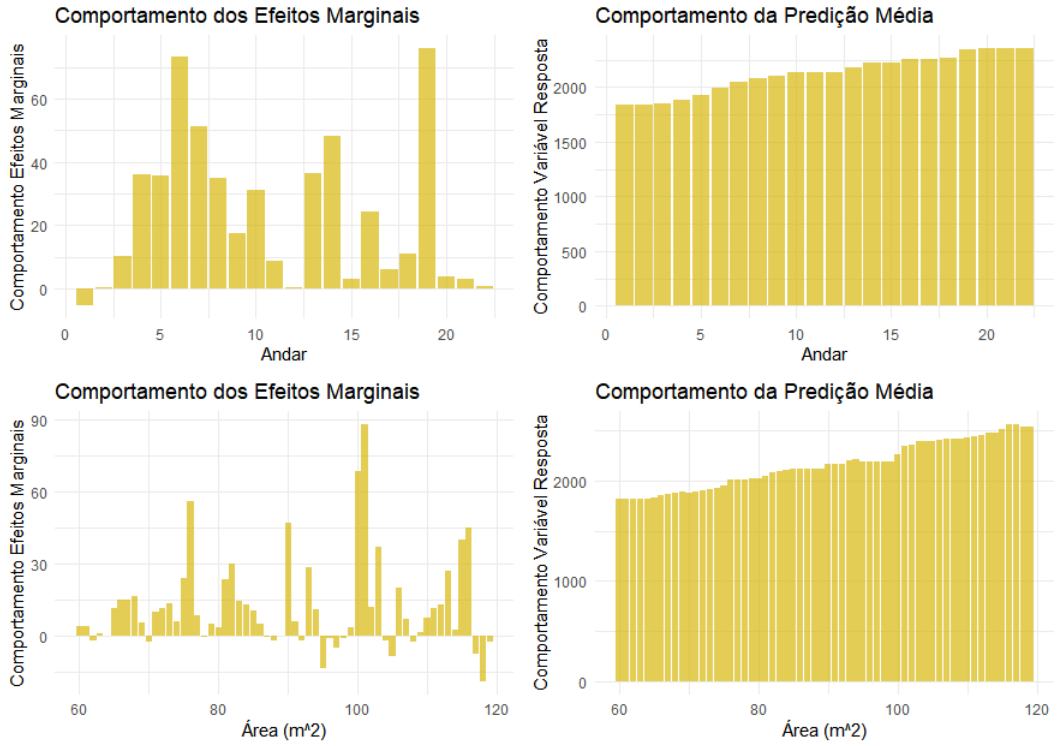
Esse tipo de efeito poderia ser aproximado com regressões pseudo-quantílicas, em que estimamos por OLS um modelo por quantil da amostra ao invés de seguir o procedimento de minimização de valor absoluto da regressão quantílica 'verdadeira'. Uma limitação dessa abordagem é que envolve fatiar a amostra. Mais quantis implicam em amostras sucessivamente menores. Essa limitação não está presente em florestas aleatórias.

Figura 16 – Cada curva contém a previsão de uma floresta aleatória distinta.



Fonte – Elaboração própria.

Figura 17 – Em florestas aleatórias a heterogeneidade dos efeitos de tratamento fica evidente.



Fonte – Elaboração própria.

Outra diferença é que modelos lineares são *globais*. Os efeitos marginais independem dos níveis de outros regressores então o salto do térreo para o primeiro andar seria tido como igual para todo apartamento, em toda cidade, de qualquer tamanho. Esse tipo de nuance não é perdida em modelos de florestas aleatórias.

Embora o exemplo seja em um contexto de regressão, o procedimento não se perde em classificação. A classe predita por uma floresta é a apenas a que recebeu mais votos das árvores individuais. Podemos ler a fração dos votos que a classe mais votada levou como uma probabilidade predita e calcular os efeitos marginais da mesma maneira que faríamos em um caso de regressão, avaliando o efeito sobre a probabilidade predita. Essa mudança para que a saída do modelo seja em probabilidade é trivial na maioria das implementações computacionais destes métodos e não leva a um aumento de complexidade da tarefa.

5 CONSIDERAÇÕES FINAIS

Este trabalho se compromete com reprodutibilidade e ciência aberta. Todo o código de estimação de modelos, preparação dos dados, tabelas, imagens e implementação do procedimento está disponível no repositório https://github.com/pedrocava/monografiaRandomForest.

Foi demonstrado como construir uma floresta aleatória partindo de primeiros princípios. Usando conceitos simples de Teoria dos Grafos, construímos uma representação matemática para processos de decisão e exploramos alguns procedimentos de escolha dessas regras a partir de dados. Tendo um processo de estimação, vimos como algums propriedades estatísticas de árvores de decisão, como sua alta variância, inibiam o uso aplicado. Para adereçar isso, agregamos várias árvores em uma floresta aleatória.

Partindo daí, foi ilustrada a Teoria Clássica de Regressão Linear. Vimos como encontrar os efeitos marginais nessa classe de modelos era muito simples, assim como era problemático recuperar essas grandezas de um modelo de floresta aleatória. Um procedimento computacionalmente simples para isso foi apresentado. Aplicamos o procedimento em um experimento controlado e em dados coletados do mundo real, para notar que não só o procedimento aproxima os mesmos efeitos marginais de um modelo linear com razoável precisão, bem como capturou nuances nos preços de alugueis que eram perdidas por modelos lineares.

Uma limitação não-resolvida da técnica é a inferência dos efeitos marginais computados. É possível distingui-los estatisticamente de zero com algum teste de hipótese? Isso equivale a estabelecer uma ponte entre as duas culturas de Breiman (2001). A técnica aqui desenvolvida pode ser estendida para aceitar testes apropriados, aproximando as duas abordagens.

Não parece um problema intratável. Se florestas aleatórias têm distribuição normal e estimamos várias, então temos dados e podemos fazer inferência se conhecermos a variância da floresta. Não é um problema trivial, mas está ao alcance da pesquisa. Essa avenida de pesquisa é provavelmente a de maior interesse ao econometrista aplicado e parte diretamente dos princípios elaborados nesta monografia.S

6 REFERÊNCIAS

ABNTEX, E. Manual de uso do pacote abntex2cite: tópicos específicos da abnt nbr 10520: 2002 e o estilo bibliográfico alfabético (sistema autor-data.[sl], 2012. *Disponível e m:< http://code. google. com/p/abntex2*. Nenhuma citação no texto.

ABNTEX, E. Modelo canônico de trabalhos acadêmicos com abntex2.[sl], 2012. *http://code. google. com/p/abntex2/>. Citado*, v. 2, p. 2. Nenhuma citação no texto.

BASU, A. On the rate of approximation in the central limit theorem for dependent random variables and random vectors. *Journal of Multivariate Analysis*, Elsevier, v. 10, n. 4, p. 565–578, 1980. Citado na página 20.

BOLLOBÁS, B. *Modern graph theory*. [S.l.]: Springer Science & Business Media, 2013. v. 184. Citado na página 12.

BOTCHKAREV, A. Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology. *arXiv preprint arXiv:1809.03006*, 2018. Citado na página 32.

BREIMAN, L. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, Institute of Mathematical Statistics, v. 16, n. 3, p. 199–231, 2001. Citado 2 vezes nas páginas 9 e 40.

BREIMAN, L. et al. *Classification and regression trees*. [S.l.]: CRC press, 1984. Citado 2 vezes nas páginas 9 e 17.

CHARTRAND, G.; ZHANG, P. *A first course in graph theory*. [S.l.]: Courier Corporation, 2013. Citado na página 12.

CLAESEN, M.; MOOR, B. D. Hyperparameter search in machine learning. *arXiv preprint arXiv:1502.02127*, 2015. Citado na página 34.

DERUMIGNY, A.; SCHMIDT-HIEBER, J. *On lower bounds for the bias-variance trade-off*. 2020. Citado na página 19.

EDISON, H.; CARCEL, H. Text data analysis using latent dirichlet allocation: an application to fomc transcripts. *Applied Economics Letters*, Taylor & Francis, p. 1–5, 2020. Citado na página 10.

FEURER, M.; HUTTER, F. Hyperparameter optimization. In: *Automated Machine Learning*. [S.l.]: Springer, Cham, 2019. p. 3–33. Citado na página 32.

FLACH, P. Performance evaluation in machine learning: The good, the bad, the ugly, and the way forward. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. [S.l.: s.n.], 2019. v. 33, p. 9808–9814. Citado na página 10.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. *The elements of statistical learning*. [S.l.]: Springer series in statistics New York, 2001. v. 1. Citado na página 9.

HASTIE, T.; TIBSHIRANI, R.; WAINWRIGHT, M. *Statistical Learning with Sparsity: the LASSO and generalizations*. [S.l.]: Chapman and Hall/CRC, 2015. Citado na página 20.

HAYASHI, F. *Econometrics*. [S.l.]: Princeton University Press, Princeton, 2000. Citado na página 21.

KUHN, M.; WICKHAM, H. *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles.* [S.l.], 2020. Disponível em: <https://www.tidymodels.org>. Citado na página 31.

KULLBACK, S.; LEIBLER, R. A. On Information and Sufficiency. *The Annals of Mathematical Statistics*, JSTOR, v. 22, n. 1, p. 79–86, 1951. Citado na página 18.

LAWRENCE, I.; LIN, K. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, JSTOR, p. 255–268, 1989. Citado na página 33.

LEPRI, B. et al. Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology*, Springer, v. 31, n. 4, p. 611–627, 2018. Citado na página 10.

LIAW, A.; WIENER, M. Classification and regression by randomforest. *R News*, v. 2, n. 3, p. 18–22, 2002. Disponível em: <https://CRAN.R-project.org/doc/Rnews/>. Citado na página 31.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, Springer, v. 5, n. 4, p. 115–133, 1943. Citado na página 9.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2020. Disponível em: <https://www.R-project.org/>. Citado na página 31.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*, 2016. Citado na página 9.

ROSENBLATT, F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, American Psychological Association, v. 65, n. 6, p. 386, 1958. Citado na página 9.

SHAHRIARI, B. et al. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, IEEE, v. 104, n. 1, p. 148–175, 2015. Citado na página 32.

STROBL, C.; MALLEY, J.; TUTZ, G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, American Psychological Association, v. 14, n. 4, p. 323, 2009. Citado na página 18.

VAPNIK, V.; CHERVONENKIS, A. *Theory of pattern recognition*. [S.l.]: Nauka, Moscow, 1974. Citado na página 9.

WRIGHT, M. N.; ZIEGLER, A. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, v. 77, n. 1, p. 1–17, 2017. Citado na página 31.