UNIVERSIDADE FEDERAL DE MINAS GERAIS
CURSO DE ESPECIALIZAÇÃO EM ESTATÍSTICA

MODELO PREDITIVO DO PREÇO DE VENDA DE APARTAMENTOS EM BELO HORIZONTE UTILIZANDO RANDOM FOREST

EVANDRO HENRIQUE MARTINS NERI

BELO HORIZONTE
2020

EVANDRO HENRIQUE MARTINS NERI

MODELO PREDITIVO DO PREÇO DE VENDA DE APARTAMENTOS EM BELO HORIZONTE UTILIZANDO RANDOM FOREST

Versão final

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Estatística da Universidade Federal de Minas Gerais, como requisito parcial para a obtenção do título de Especialista em Estatística.

Área de concentração: Estatística

Orientadora: Ilka Afonso Reis
Universidade Federal de Minas Gerais

UNIVERSIDADE FEDERAL DE MINAS GERAIS
CURSO DE ESPECIALIZAÇÃO EM ESTATÍSTICA
BELO HORIZONTE
2020

, Neri Evandro Henrique Martins

N445m Modelo preditivo do preços de venda de apartamentos em Belo Horizonte utilizando random forest [manuscrito] / Evandro Henrique Martins Neri. — 2020.
xii, 30.f. il.

Orientadora: Ilka Afonso Reis.
Monografia (especialização) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística.
Referências 29-30

1. Estatística. 2. Bens imóveis - Avaliação. 3. Mercado imobiliário. 4. Random forest. I. Reis ,Ilka Afonso. II. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatísitca .III.Título.

CDU 519.2 (043)

Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende Costa CRB 6ª Região nº 1510

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística
Programa de Pós-Graduação / Especialização
Av. Pres. Antônio Carlos, 6627 - Pampulha
31270-901 – Belo Horizonte – MG

E-mail: pgest@ufmg.br
Tel: 3409-5923 – FAX: 3409-5924

ATA DO 216ª. TRABALHO DE FIM DE CURSO DE ESPECIALIZAÇÃO EM ESTATÍSTICA DE EVANDRO HENRIQUE MARTINS NERI.

Aos quatro dias do mês de novembro de 2020, às 10:00 horas, com utilização de recursos de videoconferência a distância, reuniram-se os professores abaixo relacionados, formando a Comissão Examinadora homologada pela Comissão do Curso de Especialização em Estatística, para julgar a apresentação do trabalho de fim de curso do aluno **Evandro Henrique Martins Neri**, intitulado: "*Modelo preditivo do preço de venda de apartamentos em Belo Horizonte utilizando random forest*", como requisito para obtenção do Grau de Especialista em Estatística. Abrindo a sessão, a Presidente da Comissão, Professora Ilka Afonso Reis – Orientadora, após dar conhecimento aos presentes do teor das normas regulamentares, passou a palavra ao candidato para apresentação de seu trabalho. Seguiu-se a arguição pelos examinadores com a respectiva defesa do candidato. Após a defesa, os membros da banca examinadora reuniram-se sem a presença do candidato e do público, para julgamento e expedição do resultado final. Foi atribuída a seguinte indicação: o candidato foi considerado Aprovado condicional às modificações sugeridas pela banca examinadora no prazo de 30 dias a partir da data de hoje por unanimidade. O resultado final foi comunicado publicamente ao candidato pela Presidente da Comissão. Nada mais havendo a tratar, a Presidente encerrou a reunião e lavrou a presente Ata, que será assinada por todos os membros participantes da banca examinadora. Belo Horizonte, 04 de novembro de 2020.

Profa. Ilka Afonso Reis (Orientadora)
Departamento de Estatística / UFMG

Profa. Magda Carvalho Pires
Departamento de Estatística / UFMG

Prof. Marcos Oliveira Prates
Departamento de Estatística / UFMG

Resumo

Proprietários de imóveis podem enfrentar dificuldades no momento da venda para descobrir um preço a ser pedido. O método tradicional para obter uma estimativa deste valor é a solicitação de um parecer de um corretor imobiliário que usa a sua experiência para obter uma estimativa. Estas avaliações podem ser bastante específicas para as características e condições incomuns de imóveis, mas geram uma limitação na velocidade de entrega do resultado. Este projeto propõe a criação de um modelo com Random Forest para retornar uma estimativa útil do preço de apartamentos na cidade de Belo Horizonte, Minas Gerais. O modelo construído obteve no conjunto de teste um erro médio absoluto de 8,21% e um R^2 de 93,92%. Este resultado pode ser considerado satisfatório, permitindo recomendar esta abordagem para uma estimativa inicial imediata do preço de um apartamento.

Palavras-chave: Avaliação de imóveis. Mercado imobiliário. Random Forest.

Abstract

Real estate property owners may face difficulties at the time of selling to find a price to be asked. The traditional method for obtaining an estimate of this value is to request an opinion from a real estate agent who uses his experience to make an estimate. These assessments can be quite specific for the unusual characteristics and conditions of real estate, but they generate a limitation on the speed of delivery of the result. This project proposes the creation of a model with Random Forest to return a useful estimate of the price of apartments in the city of Belo Horizonte, Minas Gerais. The built model obtained an absolute mean error of 8.21% and an R^2 of 93.92% in the test set. This result can be considered satisfactory, allowing to recommend this approach for an immediate initial estimate of the price of an apartment.

Keywords: Real estate valuation. Real estate market. Random Forest.

Lista de Figuras

Lista de Abreviaturas e Siglas

Bagging	Bootstrap Aggregating
CART	Classification and Regression Trees
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
NRMSE	Normalized Root Mean Squared Error
RMSE	Root Mean Squared Error
SHAP	Shapley Additive Explanation
SSE	Sum of Squares Error

Sumário

1 Introdução

No cenário atual do mercado imobiliário é cada vez mais raro encontrar imobiliárias que não consideram a internet como sua principal ferramenta para obter novos clientes. O processo de digitalização da venda de imóveis aconteceu primeiramente na obtenção de novos clientes, permitindo uma maior capilaridade no que antes estava limitado em uma pequena zona de atuação. Em um momento seguinte as imobiliárias começaram a perceber que seria necessária uma maior transformação digital para acompanhar de forma escalável as novas possibilidades de crescimento.

O processo de venda de um imóvel através da intermediação de uma imobiliária consiste em quatro etapas: captação, prospecção de cliente comprador, proposta e fechamento do contrato. A definição do preço de um imóvel é o maior ponto de atrito nas três primeiras etapas do processo. Os atritos iniciam no momento da escolha do preço pedido, seguem na dificuldade de encontrar interessados para um imóvel com preço elevado e terminam em propostas muito abaixo do que proprietário inicialmente desejaria. Atuar sobre o problema desde o momento da captação pode evitar estes gargalos.

1.1 Justificativa

Um grande problema no momento da captação de um imóvel por uma imobiliária está em sugerir um preço adequado que possibilite a venda em tempo razoável. Em uma imobiliária um típico processo de captação consiste nas seguintes etapas:

1. Um proprietário entra em contato desejando anunciar seu imóvel na imobiliária.
2. Uma visita de um corretor associado é agendada para coletar informações do imóvel.
3. A avaliação do preço de venda do imóvel é realizada em uma reunião de corretores associados.
4. O resultado da avaliação é comunicado ao proprietário e ele decide qual o preço deseja pedir pelo imóvel.
5. O anúncio do imóvel é publicado.

Uma avaliação imediata de precificação pelo corretor, no ato da primeira visita, poderia substituir as reuniões e reduzir drasticamente o tempo entre a visita inicial e a publicação do anúncio. No entanto, uma avaliação pontual oferecida por um único corretor não é o padrão adotado usualmente. Avaliar pontualmente um imóvel pode gerar insatisfação do proprietário, no caso de uma sugestão de valor de venda inferior ao esperado.

Este trabalho pretende desenvolver um modelo preditivo que permita estimar o preço de venda por meio das características gerais de um imóvel e justificar esta estimativa. A ferramenta poderá ser utilizada por corretores no momento da visita ao imóvel, possibilitando ao corretor apresentar uma avaliação imediata e explicar ao proprietário o resultado obtido. O uso do modelo permite ao corretor iniciar uma discussão sobre o preço pedido pelo proprietário baseando em um valor previsto de maneira impessoal.

1.2 Organização do Trabalho

A fundamentação teórica é apresentada no Capítulo 2, explicando tecnicamente alguns métodos utilizados no trabalho. A metodologia é descrita no Capítulo 3, ressaltando os passos para desenvolver um modelo que atenda aos objetivos propostos. O Capítulo 4 apresenta os resultados obtidos com o modelo. A conclusão do trabalho é apresentada no Capítulo 5.

2 Fundamentação Teórica

Este trabalho pretende apresentar uma modelagem utilizando o método *Random Forest* e técnicas auxiliares. Portanto, é importante estabelecer os conceitos relacionados com a implementação ou que auxiliam a compreensão dos resultados.

2.1 Árvores de Decisão

O método de árvore de decisão para regressão e classificação implica em estratificar ou segmentar o espaço preditor em uma quantidade de regiões simples. Este espaço preditor pode ser entendido como um espaço formado pelas variáveis independentes e seus possíveis valores. Uma previsão é realizada tipicamente através da média ou moda das observações do conjunto de treinamento que fazem parte da região onde a nova observação se enquadrou. Este conjunto de regras de particionamento pode ser representado na forma de uma árvore, dando origem ao nome. (JAMES et al., 2013)

Este trabalho foca especificamente nas árvores de decisão para regressão, pois a previsão desejada é uma quantificação. Para a construção de árvores de regressão um dos métodos mais utilizados é o proposto por Breiman et al. (1984), intitulado *Classification and Regression Trees* (CART). "Ao contrário de muitos outros procedimentos estatísticos que foram transferidos de lápis e papel para calculadoras e depois para computadores; esse uso de árvores era impensável antes dos computadores". (BREIMAN et al., 1984)

O método CART utiliza o conjunto de dados, S, buscando cada valor distinto de cada variável independente para encontrar simultaneamente a variável e o ponto de particionamento que minimiza o total da soma dos quadrados dos erros nos grupos S_1 e S_2. Em seguida para cada um destes grupos o método repete o procedimento, sendo por este motivo também conhecido como particionamento recursivo. Esta soma dos quadrados dos erros, também denominada Sum of Squares Error (SSE) é obtida através da soma dos quadrados dentro de cada uma das partições, como fica demonstrado na Equação (1), no qual y_i é o valor real no treinamento, $\bar{y}_1$ é o valor médio da primeira partição e $\bar{y}_2$ é o valor médio da segunda partição. (KUHN; JOHNSON, 2013)

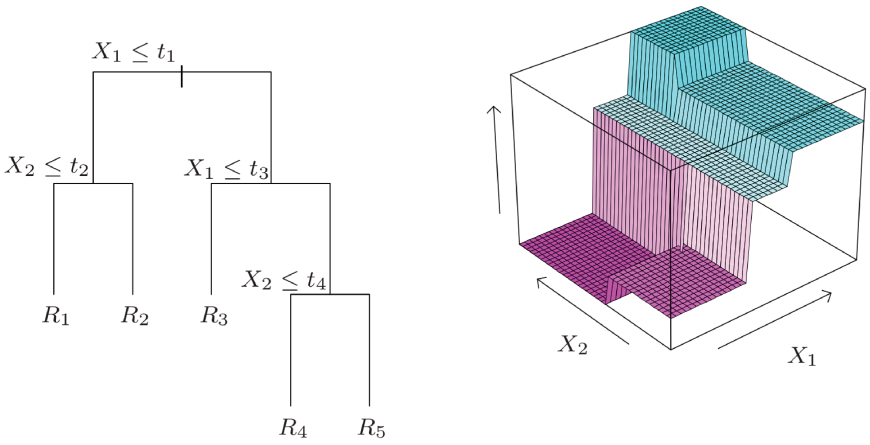
$$\mathrm{SSE} = \sum_{i \in S_1} (y_i - \bar{y}_1)^2 + \sum_{i \in S_2} (y_i - \bar{y}_2)^2 \tag{1}$$

O particionamento obtido pela construção de uma árvore pode ser melhor compreendido pela análise da Figura 1. No lado esquerdo é apresentado uma árvore construída de um

espaço bidimensional de variáveis X_1 e X_2. Os pontos de decisão $t_1, ..., t_4$, orientarão em qual das regiões $R_1, ..., R_5$ uma nova observação será direcionada, resultando em uma previsão da média do conjunto de treinamento daquela região. (HASTIE; TIBSHIRANI; FRIEDMAN, 2009)

Segundo Breiman et al. (1984), "a árvore pode ser pensada como estimativa de histograma da superfície de regressão". No lado direito da Figura 1 essa descrição fica mais clara através da visualização de perspectiva da superfície correspondente à árvore.

Figura 1 – Particionamento de uma árvore de decisão



Fonte: Hastie, Tibshirani e Friedman (2009)

2.2 Bagging

A construção de árvores de decisão pode gerar modelos instáveis, onde uma pequena mudança no conjunto de dados de treinamento pode provocar grandes mudanças na previsão. Para melhorar a estabilidade destes modelos Breiman (1996) propôs um procedimento chamado *Bootstrap Aggregating* (bagging).

O *Bagging* é um procedimento para redução de variância de métodos de aprendizado estatístico, muito útil no contexto de árvores de decisão. Considerando um conjunto de n observações independentes, $Z_1, ..., Z_n$, tendo cada uma variância σ^2, a variância da média $\bar{Z}$ das observações é dada por σ^2/n. Portanto, um método para reduzir a variância e aumentar a acurácia da predição consiste em obter muitos conjuntos de treinamento, através de amostragem repetida do conjunto de treinamento original, construir distintos modelos e obter a médias das previsões. Calcularíamos a média de diversas previsões $\hat{f}^1(x)$, $\hat{f}^2(x), ..., \hat{f}^B(x)$ de modelos utilizando B conjuntos de treinamento, visando obter um único modelo de baixa variância. A Equação (2) apresenta a média dos B modelos, sendo

$\hat{f}^{*b}(x)$ a estimativa gerada por um modelo b individual em uma das amostras do conjunto original de dados. (JAMES et al., 2013)

$$\hat{f}_{\text{bagging}}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{*b}(x) \tag{2}$$

2.3 Random Forest

Árvores de decisão frequentemente não podem crescer em complexidade sem que, em contrapartida, possa ocorrer redução de poder de generalização para novos dados. Ho (1995) propõe um método denominado *Random Decision Forests* para remover a limitação de complexidade das árvores no treinamento. A forma sugerida para construção dos modelos permite que a complexidade das árvores seja aumentada arbitrariamente, sem que isso resulte na perda de desempenho em dados ainda não vistos. O método propõe construir diversas árvores em um subespaço selecionado do espaço de variáveis preditoras, para que estas generalizem o resultado de formas complementares. Por fim, o resultado agregado destas árvores construídas sem limitação de profundidade aumenta o desempenho do modelo sem que isso resulte na perda de poder de generalização.

O método *bagging* para gerar múltiplas versões de um mesmo modelo foi apresentado por Breiman (1996), resultando em demonstrações de ganhos de acurácia. Dentro os resultados do estudo ficou claro que o desempenho de árvores de decisão com agregação de *bootstrap* foi significativo. Este método proposto se tornou um caminho promissor para tentar melhorar métodos existentes, pois sua implementação é relativamente simples.

O método *Random Forest* utilizado neste trabalho foi proposto por Breiman (2001). A construção do modelo é feita através da geração de amostras *bootstrap* e utilizando um subconjunto aleatório de variáveis candidatas para a divisão em cada nó. A forma da amostragem do conjunto de dados é o proposto em Breiman (1996). A seleção do subconjunto de variáveis para reduzir a correlação das árvores é similar ao proposto em Ho (1995), se diferenciando apenas por gerar subconjuntos no nível dos nós e não para toda a árvore.

O Algoritmo 1 explica as etapas para a construção de uma *Random Forest*, sendo B o número de total de árvores, n o número de elementos do nó naquela iteração e n_{min} o mínimo de elementos para que um nó possa continuar sendo dividido. A iteração externa obtém as amostras *bootstrap*, enquanto a iteração interna seleciona aleatoriamente as variáveis que serão candidatas para a divisão.

Para realizar uma previsão para uma nova observação x empregamos a Equações (3) e (4), sendo adequadas respectivamente para regressão e classificação. A Equação (3), no qual $T_b(x)$ é a previsão da árvore b para a observação x, retorna a média dos resultados das B

Algoritmo 1: Random Forest

for $b = 1...B$ **do**
  Obtenha uma amostra bootstrap Z^* de tamanho N do conjunto de treinamento;
  while $n < n_{min}$ **do**
    Inicie a construção de uma árvore T_b na amostra obtida;
    Selecione aleatoriamente m dentre todas as p variáveis;
    Escolha a melhor relação de variável e ponto de divisão dentre m;
    Divida o nó em dois nós filhos;
  end
end
Retorne o conjunto de árvores $\{T_b\}_1^B$;

Fonte: Hastie, Tibshirani e Friedman (2009)

árvores de regressão. A Equação (4), no qual $\hat{C}_b(x)$ é a classificação da árvore b para a observação x, resultando na moda das previsões, portanto, a classe mais prevista das B árvores de classificação. (HASTIE; TIBSHIRANI; FRIEDMAN, 2009)

$$\hat{f}_{\mathrm{rf}}^B(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \tag{3}$$

$$\hat{C}_{\mathrm{rf}}^B(x) = \text{ moda } \left\{\hat{C}_b(x)\right\}_1^B \tag{4}$$

A construção de um modelo *Random Forest* é relativamente simples, não existe necessidade de tratamento especial de variáveis qualitativas. O pequeno número de hiperparâmetros, que são os parâmetros definidos previamente para controlar o treinamento, facilita encontrar valores que proporcionam bom desempenho. Etapas mais complexas estavam relacionadas com pré-processamentos nas fases de tratamento, imputação e seleção de variáveis. É necessário definir os valores para o número de árvores, o número de variáveis aleatoriamente selecionadas em cada particionamento e o tamanho mínimo do nó. Segundo Wright e König (2019), variáveis qualitativas ordinais podem ser tratadas do mesmo modo que preditores numéricos, produzindo o mesmo resultado da abordagem que trata estes valores como nominais, porém com menor complexidade computacional.

Hastie, Tibshirani e Friedman (2009) aponta que os valores padrão para o número de variáveis aleatoriamente selecionadas e para o tamanho mínimo do nó podem não ser ideais, devendo ser tratados como parâmetros a serem ajustados. A estratégia para obter valores satisfatórios utiliza a busca em grade. Desta maneira todas as combinações possíveis dos valores determinados como razoáveis para cada hiperparâmetro são testadas através de validação cruzada. O melhor resultado é adotado para um modelo final.

O tamanho mínimo do nó está relacionado com o tamanho da árvore, sendo árvores menores resultado de valores maiores para este hiperparâmetro.

O número de árvores também pode ser visto como um hiperparâmetro, mas foi tratado de forma diferente. Probst e Boulesteix (2017) argumenta que um valor suficientemente alto deve ser adotado sem que isso prejudique os resultados, levando em consideração a capacidade computacional disponível. Segundo Genuer, Poggi e Tuleau (2008) um número menor de árvores pode obter um resultado próximo de um conjunto maior de árvores, porém o número maior de árvores aumenta a estabilidade das importâncias das variáveis.

Para regressão com *Random Forest* podemos utilizar diversas métricas. A Equação (5) demonstra a mais popular métrica, raiz do erro quadrático médio, conhecida como Root Mean Squared Error (RMSE), que tem uma escala dependente dos dados e maior penalização para maiores erros. Outra métrica dependente da escala dos dados é o erro médio absoluto da Equação (6), conhecido também como Mean Absolute Error (MAE), divergindo do RMSE por não penalizar os maiores erros de forma diferenciada.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{5}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}| \tag{6}$$

Métricas que não dependem da escala dos dados são bastante interpretáveis. A Equação (7) apresenta a métrica de erro médio absoluto percentual, conhecida como Mean Absolute Percentage Error (MAPE), que ajuda a compreender um erro percentual que deve ser esperado em média. A Equação (8) demonstra o coeficiente de determinação, também conhecido como R^2, podendo ser compreendido como o percentual da variância da variável dependente que o modelo consegue explicar utilizando as variáveis independentes.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \cdot 100 \tag{7}$$

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2} \tag{8}$$

2.4 MissForest

Dados faltantes são um problema muito comum que impede a aplicação de diversos tipos de análise. Na visão frequentista da Estatística, conclusões de um estudo não deveriam

depender de uma amostra específica, pois resultados próximos deveriam ser obtidos para uma amostra diferente. O conceito de imputação é baseado neste princípio, em que cada observação aleatoriamente escolhida pode ser substituida por uma outra aleatoriamente selecionada da população sem o comprometimento das conclusões. Existem muitos métodos para lidar com dados faltantes, os mais comuns e simples são a análise de casos completos, o método de indicador de faltante e a imputação da média geral. Estes métodos podem gerar análises ineficientes, podendo ser mais adequado realizar a imputação de um valor em relação às demais características conhecidas da observação. (DONDERS et al., 2006)

O algoritmo *MissForest* para imputação apresenta simulaneamente várias propriedades desejáveis. É capaz de lidar com dados contínuos e categóricos. Não são necessários ajustes de parâmetros ou pressuposição sobre a distribuição. A utilização de *Random Forest* permite, durante a construção de cada árvore, estimar o erro de imputação através da previsão na parcela das observações fora da amostra. (STEKHOVEN; BUHLMANN, 2011)

O método é apresentado no Algoritmo 2. Inicialmente uma estimativa inicial é gerada para os valores faltantes em X, usando a média ou outro método de imputação. Em seguida as variáveis $X_s, s = 1, ..., p$ são ordenadas crescentemente de acordo com o número de valores faltantes. Para cada varíavel X_s, os valores faltantes são imputados primeiramente por uma *Random Forest* gerada com variável resposta $y_{obs}^{(s)}$, que são os valores observados da variável que terá valores faltantes imputados. As variáveis preditoras $x_{obs}^{(s)}$ são constituídas por todas as demais características. O modelo gerado é utilizadas para prever os valores faltantes $y_{mis}^{(s)}$ através das variáveis preditoras $x_{mis}^{(s)}$. O processo de imputação é repetido até que a diferença da nova matriz imputada para a anterior aumente pela primeira vez. (STEKHOVEN; BUHLMANN, 2011)

Segundo Stekhoven e Buhlmann (2011) é possível avaliar a eficiência da imputação utilizando a versão normalizada da raiz do erro quadrático médio, conhecida também como Normalized Root Mean Squared Error (NRMSE). A Equação (9) demonstra que o desempenho é avaliado através dos erros nas observações fora da amostra de treinamento como conjunto de teste durante a construção das árvores da *Random Forest*. A média do quadrado deste erro é normalizado pela variância dos valores reais da variável nas observações não faltantes, e por fim a raiz quadrada deste valor é o que desejamos. Valores próximos de 0 indicam um bom desempenho, valores próximos de 1 indicam um resultado que não é

Algoritmo 2: MissForest

Gere uma estimativa inicial para os valores faltantes;
$k \leftarrow$ vetor dos índices das colunas ordenados de forma crescente pela quantidade de valores faltantes;
while *not* γ **do**
    $X_{old}^{imp} \leftarrow$ matriz imputada anteriormente;
    for s *in* k **do**
        Construa uma Random Forest: $y_{obs}^{(s)} \sim x_{obs}^{(s)}$;
        Preveja $y_{mis}^{(s)}$ usando $x_{mis}^{(s)}$;
        $X_{new}^{imp} \leftarrow$ matriz imputada usando o $y_{mis}^{(s)}$ predito;
    end
    $\gamma \leftarrow$ indicação de aumento de diferença da nova matriz imputada para a anterior;
end
Retorne a matriz imputada x^{imp};

Fonte: Stekhoven e Buhlmann (2011)

melhor que utilizar uma simples imputação com a média dos valores observados da variável.

$$\mathrm{NRMSE} = \sqrt{\frac{1}{\mathrm{var}\left(X^{\mathrm{true}}\right)}\left(\frac{1}{n}\sum_{i=1}^{n}\left(X^{\mathrm{true}} - X^{\mathrm{imp}}\right)^2\right)} \tag{9}$$

2.5 Boruta

A seleção de variáveis é uma etapa de decisão sobre a relevência de cada variável. O primeiro desafio está na própria definição de relevância. Para fixar este conceito serão adotados dois níveis de relevância: forte e fraca. Uma variável X é fortemente relevante se a remoção isolada de X resultar na deterioração do desempenho do modelo. Uma variável X é fracamente relevante se não for fortemente relevante e existir um subconjunto de variáveis, S, tal que o desempenho em S é pior que em $S \cup X$. Uma variável é irrelevante se não for forte ou fracamente relevante. (KOHAVI; JOHN, 1997)

O problema de seleção de variáveis se divide em duas direções: mínimo-ótimo e todo-relevante. O problema do mínimo-ótimo consiste em encontrar todas as variáveis fortemente relevantes e qualquer subconjunto de variáveis fracamente relevantes que não possua apenas informações redundantes. O problema do todo-relevante, que é o seguido neste trabalho, identifica todas as variáveis forte e fracamente relevantes, excluindo apenas as irrelevantes. (NILSSON et al., 2007)

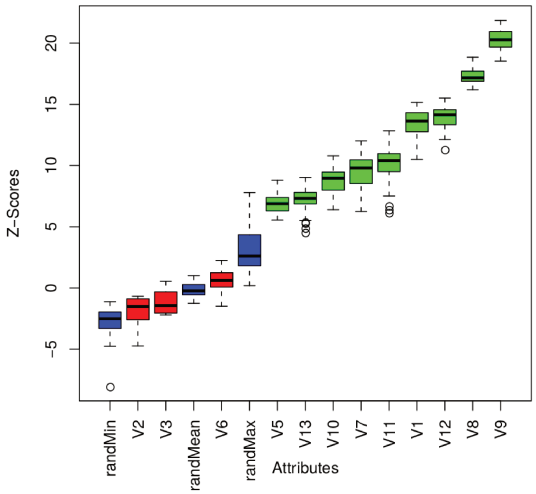
Duas dificuldades surgem para encontrar variáveis fracamente relevantes. Primeiramente a possibilidade delas serem ofuscadas por outras variáveis. Em segundo lugar, existe a necessidade de diferenciar variáveis-ruído que possam parecer relevantes simplesmente

pelo acaso. As características de uma *Random Forest* são adequadas para esta tarefa, pois durante a sua construção cada variável possui inúmeras chances de ser incluída no modelo. As fracamente relevantes também têm sua chance, portanto deixando visíveis as contribuições mesmo se utilizadas em um pequeno número de árvores. (KURSA; RUDNICKI, 2011)

Kursa e Rudnicki (2010) propõe um algoritmo, denominado Boruta, que avalia a relevância de uma variável comparando a importância obtida através de uma *Random Forest* para as variáveis originais com a importância para as variáveis aleatórias adicionadas artificialmente. Um modelo *Random Forest* é treinado em um conjunto de dados estendido com variáveis de contraste aleatório. Estas são obtidas pela mistura aleatória de valores originais entre objetos. Para cada variável selecionada a importância deve ser maior que a importância máxima alcançada por qualquer variável de contraste. De modo a obter resultados estatisticamente significantes, as etapas anteriores são repetidas inúmeras vezes com variáveis de contraste que foram geradas de maneira independente em cada iteração. A medida de importância é obtida pela média das perdas de acurácia em todas as árvores causadas pela mistura aleatória. Este valor é dividido pelo desvio padrão destas perdas, gerando o escore Z que será utilizado para comparação com as variáveis contraste.

A Figura 2 mostra o resultado obtido para um exemplo com 12 variáveis, sendo 9 com importância maior que a importância máxima alcançada aleatoriamente. Outras 3 não foram consideradas significativas, pois obtiveram importância menor que o máximo esperado apenas pelo acaso. Os escore Z sozinhos não fornecem informação suficiente para a confirmação ou rejeição de uma variável, por este motivo as variáveis contraste servem de referência externa para as decisões. (KURSA; RUDNICKI, 2010).

Figura 2 – Resultado de exemplo do Boruta, sendo variáveis confirmadas em verde, rejeitadas em vermelho e contraste em azul



Fonte: Kursa e Rudnicki (2010)

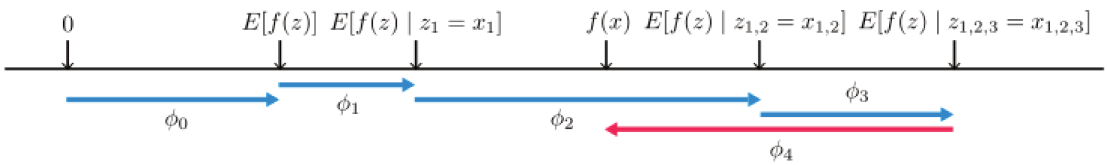
2.6 SHAP

Com os avanços da ciência e tecnologia na área de aprendizado de máquina, modelos complexos e menos interpretáveis passaram a ser mais utilizados. Essa mudança de direção torna difícil a tarefa de compreender o modelo através do seu funcionamento interno. O aspecto do papel dos humanos na utilização dos modelos como ferramentas, muitas vezes é pouco considerado, desprezando a ideia de que, se eles não confiarem no modelo, eles não o utilizarão. Para mitigar este problema a compreensão do modelo do ponto de vista dos interessados deve focar nos conceitos de confiança em uma previsão e confiança no modelo implantado. (RIBEIRO; SINGH; GUESTRIN, 2016)

Para solucionar este problema muitos métodos foram sugeridos. O método apresentado é denominado *Shapley Additive Explanations* (SHAP), e associa um valor de importância para uma variável em uma previsão em particular. A abordagem aditiva do *SHAP* é comum a vários destes métodos, mas através de resultados teóricos é possível demonstrar que este é o único da classe que atende todas as propriedades desejáveis. De forma simultânea estas propriedades somente podem ser satisfeitas através da utilização de *Shapley values*. (LUNDBERG; LEE, 2017)

Na Figura 3 fica demonstrado como partir do valor base $E[f(z)]$, que é a expectativa da previsão se não soubéssemos nenhuma das características da observação, até a previsão atual $f(x)$, considerando apenas uma única ordenação dentre as possíveis. Uma previsão para todas entradas nulas, exceto as variáveis dos índices em uma lista S, é definida por $f(z_S)$, sendo equivalente a $f(z)$ caso S seja vazio. Porém nem todos os modelos podem lidar com entradas nulas, então uma aproximação é feita de $f(z_S)$ por $E[f(z)|z_S]$. Como exemplo, $E[f(z)|z_{1,2} = x_{1,2}]$ representa a expectativa da previsão quanto todas as variáveis independentes estivessem nulas, exceto as variáveis x_1 e x_2. Os valores ϕ_i gerados são resultado da variação da previsão causada pela inclusão sequencial de novos índices de variáveis em S. Quando o modelo não é linear ou as variáveis não são independentes a ordem em que as variáveis são adicionadas importa. Os valores *SHAP* atribuem a cada variável a mudança na variação da mudança da previsão condicionada naquela variável através da média de ϕ_i em todas as ordenações posíveis. (LUNDBERG; LEE, 2017)

Figura 3 – Valores SHAP atribuindo variações de previsão por variável



Fonte: Lundberg e Lee (2017)

Na Equação (10) é apresentado o método para calcular os valores ϕ_i. O conjunto de todas as variáveis independentes é representado por x' e um possível subconjunto destas variáveis representado por z'. Uma versão simplificada das variáveis é denominada M e obtida através de uma função de mapeamento aplicada nas entradas x'. O subconjunto na ausência de uma variável específica de índice i é representado por $z' \backslash i$. Em um nível mais alto, podemos compreender o termo $f_x\left(z'\right) - f_x\left(z' \backslash i\right)$ como a diferença da previsão um modelo treinado com a presença da variável pelo resultado de outro modelo no qual a variável não foi incluída. O termo $\frac{|z'|!(M-|z'|-1)!}{M!}$ tem como objetivo calcular a média ponderada de acordo com os pesos de todas as possíveis diferenças. (LUNDBERG; LEE, 2017)

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!\,(M - |z'| - 1)!}{M!} \left[f_x\left(z'\right) - f_x\left(z' \backslash i\right)\right] \tag{10}$$

3 Metodologia

Neste capítulo são definidos os procedimentos para a construção de um modelo de previsão de preços de apartamentos na cidade de Belo Horizonte. As etapas apresentadas abrangem as técnicas e decisões, partindo da obtenção dos dados necessários até a avaliação dos resultados.

Este capítulo está dividido em cinco seções. Na Seção 3.1 discutimos os materiais utilizados, além do tratamento necessário para sua utilização. Na Seção 3.2 foi abordado o problema de dados faltantes no conjunto de dados. A seleção de variáveis para o modelo foi abordada na Seção 3.3. Na Seção 3.4 foi discutida a construção do modelo. Por fim, na Seção 3.5 foram apresentados os principais *softwares* utilizados.

3.1 Coleta e Tratamento de Dados

Os dados utilizados neste projeto foram disponibilizados pela imobiliária Casa Mineira Imóveis. Os critérios de inclusão e exclusão de imóveis neste trabalho são apresentados no Quadro 1. Esta seleção prévia visa reduzir o número de imóveis com algumas características menos frequentes no conjunto de dados. Portanto, a validade externa aqui almejada está restrita a este mesmo perfil estabelecido pelos critérios.

Quadro 1 – Critérios de inclusão e exclusão.

Critérios de inclusão	**Critérios de exclusão**
Localizado em Belo Horizonte	Preço maior que R$ 1.500.000
Anunciado no período 2017-2020	Apartamento é do tipo cobertura
É um apartamento	Apartamento possui área privativa

Estão disponíveis no conjunto original de dados 131 variáveis independentes e uma variável dependente. No Quadro 2 vemos a variável dependente **preço** e 17 das principais variáveis independentes. A maioria das variáveis são autoexplicativas, as duas variáveis menos triviais são as coordenadas de latitude e longitude, que são representadas pelos graus decimais, portanto quantitativas contínuas. A posição do imóvel é qualitativa nominal e possui 4 categorias sobre o posicionamento do imóvel em relação à rua. As demais principais variáveis são quantitativas discretas, representando com um número inteiro a quantidade para aquela variável.

Além das principais variáveis, o conjunto original de dados conta também com 114 variáveis divididas em 6 grupos de características qualitativas binárias relacionadas ao imóvel ou

Quadro 2 – Lista das principais variáveis.

Tipo	Subtipo	Variável
Quantitativa	Discreta	Preço (R$)
		Área (m^2)
		Quartos
		Suítes
		Semissuítes
		Banheiros
		Lavabos
		Vagas
		Vagas cobertas
		Salas
		Elevadores
		Número de andares
		Unidades por andar
	Contínua	Latitude (graus decimais)
		Longitude (graus decimais)
Qualitativa	Ordinal	Ano de construção
		Andar
	Nominal	Posição

ao edifício. O Quadro 3 lista os grupos, a quantidade de variáveis e um exemplo de uma variável da classe.

Quadro 3 – Lista das variáveis qualitativas nominais binárias.

Grupo	Descrição	Exemplo
Do imóvel	27 características do imóvel	Cozinha americana
Do edifício	43 características do edifício	Academia
Armários	7 locais para armário	Armário nos quartos
Pisos	16 tipos de piso	Piso de porcelanato
Revestimentos	14 tipos de revestimento	Revestimento de mármore
Esquadrias	7 tipos de esquadria	Esquadria de alumínio

A maior diferença destas características para as que foram denominadas principais reside no fato destes dados serem esparsos. Portanto, a indicação da ausência da característica nas variáveis binárias nos imóveis é muito mais frequente que a indicação de sua presença.

Ao final do tratamento dos dados eles foram separados em dois conjuntos, sendo um de

treinamento e outro de teste. Para o conjunto de treinamento 90% das observações foram separadas e utilizadas durante todo o processo de construção do modelo. Os 10% restantes foram separados no conjunto de teste, que somente foi utilizado para analisar os resultados do comportamento do modelo para previsões futuras.

3.2 Dados Faltantes

Os métodos selecionados neste trabalho para construção do modelo não são capazes de lidar com dados faltantes. Por este motivo, foi realizado um processo de imputação dos dados através do algoritmo *MissForest*. Esta escolha se justificou pelas propriedades descritas na Seção 2.4. A possibilidade de com apenas um único algoritmo poder lidar com tipos diversos de dados, bem como não ter que realizar pressuposições sobre suas distribuições são qualidades muito convenientes. Com o objetivo de avaliar os resultados da imputação, a métrica NRMSE foi utilizada, sendo desejáveis valores menores que um e idealmente próximos de zero.

3.3 Seleção de Variáveis

Ao contrário de uma única árvore de classificação, a combinação de árvores construídas utilizando variáveis selecionadas aleatoriamente como nas *Random Forests* pode produzir uma acurácia aperfeiçoada. Por este motivo selecionamos as variáveis por uma abordagem todo-relevante, no qual basta que a variável seja classificada como significante para que seja incluída no modelo, mesmo que a princípio não existam evidências do tamanho da influência desta variável.

A seleção de variáveis pelo método todo-relevante utilizando *Random Forest* foi apresentado na Seção 2.5, através do algoritmo *Boruta*. Este método foi utilizado para obter como resultado um conjunto de variáveis que possuam relação significativa com a variável dependente. Portanto, foram descartadas apenas as variáveis que obtiveram importância em faixas significativamente inferiores ao máximo de importância de variáveis-ruído geradas durante o processo.

3.4 Treinamento do Modelo

A escolha do método *Random Forest* em detrimento de um modelo mais simples como Regressão Linear foi baseada em uma restrição do conjunto de dados. Seria complexo utilizar duas das variáveis que intuitivamente podem ser consideradas muito relevantes para um imóvel: latitude e longitude. A combinação dos valores destas duas variáveis indica a localização geográfica do imóvel, e não possuem padrão linear que permita sua utilização

sem que seja necessário transformações delicadas. Além deste problema limitador, temos também propriedades no método *Random Forest* que o tornam a escolha padrão para lidar com conjuntos de dados complexos, sem que isso obrigue a análise de pressuposições e transformações de variáveis não quantitativas.

Construir um modelo *Random Forest* após a realização de todas as etapas prévias nos dados se mostra bastante simples. Resta apenas a definição dos hiperparâmetros que serão utilizados para controlar o treinamento do modelo e que devem ser definidos previamente. Eles foram definidos em intervalos, porque é possível testar de forma exaustiva a melhor combinação possível.

Foram testados para o tamanho mínimo do nó os valores 1 e 9, além do valor padrão 5. O hiperparâmetro m, que define número de variáveis aleatoriamente selecionadas em cada etapa de particionamento, foi testado com valores de p/6 até p em incrementos de p/6, sendo p o número de variáveis independentes. Para $m = p$ temos um cenário equivalente a um *bagging* de árvores de decisão. Para m igual a 1, todas as partições seriam realizadas com uma variável selecionada completamente ao acaso, portanto não foram considerados razoáveis valores muito baixos.

O método utilizado para a comparação foi a validação cruzada com 10 subconjuntos. Kohavi (1995) demonstra que esta quantidade de subconjuntos se mostrou adequada em múltiplos cenários testados para seleção de um melhor modelo dentre um conjunto de modelos.

Após todas as etapas anteriores concluídas o modelo foi construído com o conjunto de dados de treinamento, sem dados faltantes devido ao seu preenchimento através da imputação e utilizando somente as variáveis selecionadas. Para todos hiperparâmetros considerados a combinação que obteve o menor erro médio foi utilizada para o treinamento que resultou no modelo proposto. A principal avaliação dos resultados foi feita através da análise das métricas RMSE, MAE, MAPE e R^2.

3.5 Softwares Utilizados

Para o desenvolvimento deste trabalho foi empregada a linguagem *R*. O treinamento do modelo foi controlado pelo pacote *Caret*, permitindo comparar as combinações de hiperparâmetro utilizando validação cruzada. A construção do modelo *Random Forest* foi realizada através do pacote *Ranger*, que é mais adequado para conjuntos de dados com maior volume. As etapas de imputação e seleção de variáveis foram realizadas utilizando as implementações originais respectivamente dos algoritmos *MissForest* e *Boruta*.

4 Resultados

Neste capítulo são apresentados os resultados obtidos no conjunto de dados do trabalho. O conjunto de treino foi utilizado até a etapa de construção do modelo final. Após esta etapa utilizamos o conjunto de teste para avaliar o modelo gerado com estes dados que não tinham sido expostos durante sua construção.

Este capítulo está dividido em sete seções. Na Seção 4.1 realizamos uma breve análise exploratória. Imputamos dados faltantes na Seção 4.2. As variáveis significativas foram selecionadas na Seção 4.3. Na Seção 4.4 o modelo final foi gerado. A avaliação das métricas é feita na Seção 4.5. Analisamos a importância das variáveis para o modelo na Seção 4.6. Por fim na Seção 4.7 explicamos uma previsão individual.

4.1 Análise Exploratória

Após a seleção pelos critérios definidos no Quadro 1 foi obtido um conjunto com 18.796 imóveis. Com a separação 90% para treino e 10% para teste, foi gerado um conjunto de treinamento de tamanho 16.919 e um conjunto de teste de tamanho 1.877.

Na Tabela 1 são apresentadas algumas estatísticas descritivas a respeito de algumas das principais variáveis do conjunto de dados de treinamento. O preço médio dos apartamentos é R$ 484.691, o valor da mediana de R$ 400.000 é menor, indicando uma assimetria positiva com cauda mais longa à direita. O menor preço de um apartamento anunciado é de R$ 150.000 e o maior é de R$ 1.495.741, limitado pelos critérios de inclusão. A área tem como mediana 86 m^2, o número de quartos tem mediana 3, o número de banheiros e vagas tem mediana 2.

Tabela 1 – Descritivo de algumas das principais variáveis.

	Média	Desvio Padrão	Mínimo	Quartil 1	Mediana	Quartil 3	Máximo
Preço	484.691	262.314	150.000	290.000	400.000	600.000	1.495.741
Área	93,1	36	25	68	86	110	300
Quartos	2,88	0,74	1	2	3	3	5
Suítes	0,85	0,54	0	1	1	1	4
Banheiros	1,95	0,7	1	2	2	2	5
Vagas	1,68	0,78	0	1	2	2	5
Elevadores	1,09	0,96	0	0	1	2	8

As variáveis categóricas ordinais são descritas pelas suas frequências relativas e acumuladas, após agrupamento. Porém eles foram utilizados no modelo como numéricos originais, pois o método é mais eficiente computacionalmente nesta abordagem.

Analisando a Tabela 2, sobre o ano de construção dos apartamentos, é possível perceber que 52,68% dos apartamentos à venda foram construídos entre os anos de 2000 e 2020. Somente 10,28% dos apartamentos foram construídos antes de 1980.

Tabela 2 – Descritivo da variável ano de construção.

Período	Frequência relativa	Frequência acumulada
2010-2020	31,53%	31,53%
2000-2009	21,15%	52,68%
1990-1999	18,12%	70,8%
1980-1989	18,92%	89,72%
1950-1979	10,28%	100%

Observando a Tabela 3, que apresenta o andar no qual os imóveis estão situados, percebe-se que 71,98% dos apartamentos à venda estão localizados até o quinto andar e 92,44% até o décimo andar.

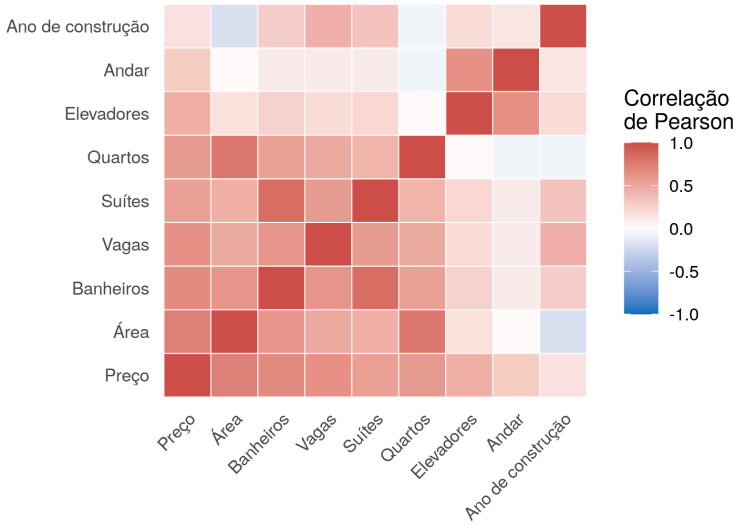
Tabela 3 – Descritivo da variável andar.

Andar	Frequência relativa	Frequência acumulada
1-5	71,98%	71,98%
6-10	20,46%	92,44%
11-33	7,56%	100%

O gráfico de correlações apresentado na Figura 4 indica que as variáveis de área, banheiros, vagas, suítes e quartos têm os maiores valores de correlação de Pearson com a variável preço, sendo todas acima de 0,5. Dentre as demais variáveis é possível perceber uma maior correlação entre andar do apartamento com o número de elevadores, banheiros com suítes e da área com o número de quartos. Analisando estas variáveis utilizando a correlação de Spearman, são obtidos valores ligeiramente superiores. Esta similaridade indica que relações monotônicas não lineares não são tão eminentes, caso contrário, estes valores seriam consideravelmente superiores aos obtidos na correlação de Pearson.

Na Figura 5 são apresentados os gráficos de dispersão de quatro das variáveis mais correlacionadas com a variável preço. A variável área aparenta ter uma relação que se

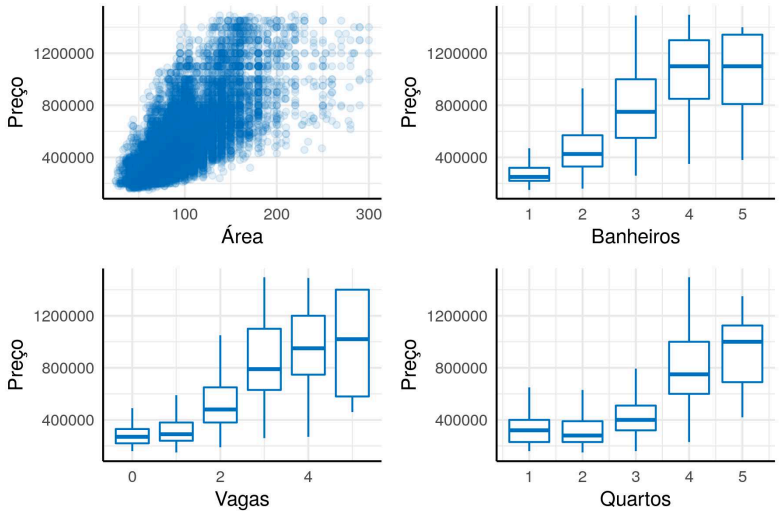
Figura 4 – Correlações das variáveis



Fonte: Elaborado pelo autor

aproxima da linear e percebe-se aumento de variância conforme ocorre aumento da área. A variável banheiro aparenta tem uma relação linear, também com aumento da variância em paralelo com o aumento do número de banheiros. As variáveis vagas e quartos possuem aumentos de preço mais visíveis nos maiores valores.

Figura 5 – Dispersão do preço em relação a algumas das principais variáveis



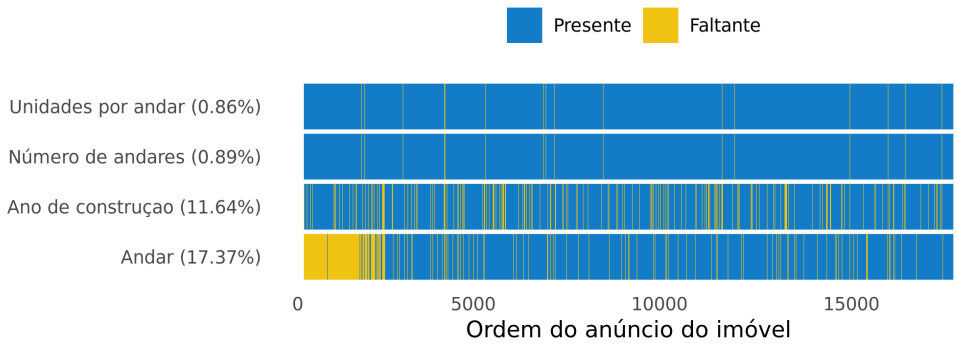
Fonte: Elaborado pelo autor

4.2 Imputação de Dados Faltantes

Na Figura 6 são listadas as 4 variáveis que possuem dados faltantes, variando entre 0,86% ocorrências até 17,37% dos casos. As linhas verticais em amarelo representam dados

faltantes e as linhas em azul representam os dados presentes. As observações estão alinhadas das mais antigas na esquerda, para as mais recentes na direita, permitindo visualizar que registros mais antigos possuem maior frequência da variável andar faltante.

Figura 6 – Ocorrências de dados faltantes



Fonte: Elaborado pelo autor

A simples utilização exclusiva de casos completos, removendo qualquer observação com alguma valor faltante, reduziria em cerca de 26,07% a quantidade de imóveis disponíveis para o treinamento. Por este motivo o caminho tomado foi a imputação dos dados faltantes.

A Tabela 4 mostra que os menores erros estimados de imputação ocorreram nas variáveis de números de andares e ano de construção do edifício, enquanto os maiores erros ocorreram nas variáveis referentes ao andar do imóvel e número de unidades por andar no edifício. Porém este cenário com erros concentrados na faixa entre 0,2 e 0,32 é relativamente melhor do que aquele que teríamos simplesmente com a imputação da média, conforme exposto na Seção 3.2.

Tabela 4 – Erros estimados de imputação.

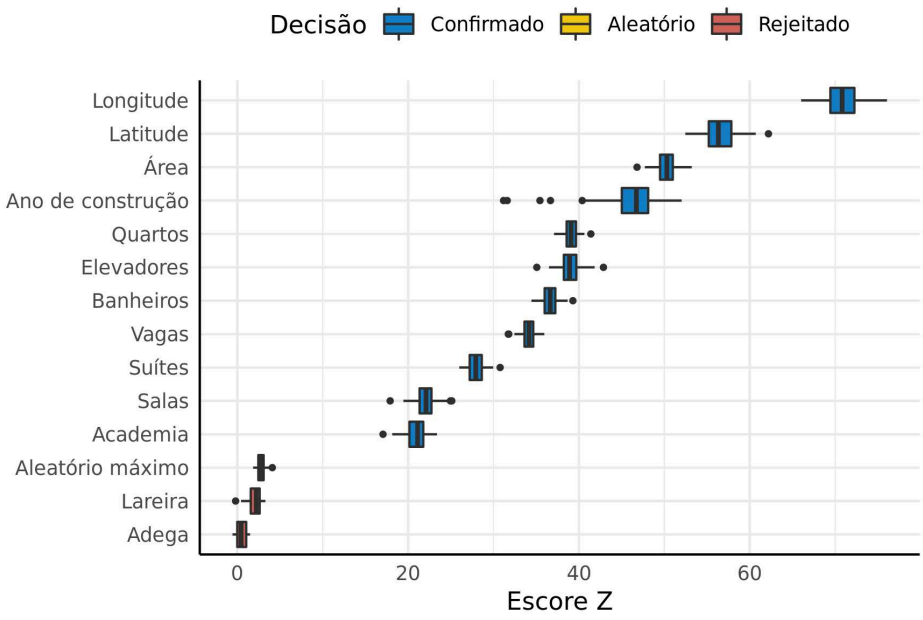
Variável	NRMSE
Número de andares	0,2
Ano de construção	0,21
Andar	0,31
Unidades por andar	0,32

4.3 Variáveis Significativas

Na Figura 7 podemos observar os resultados para 10 principais variáveis e uma amostra de 3 das 114 variáveis qualitativas binárias. Visualmente percebemos que as variáveis

principais em geral obtém escores Z altos. Algumas variáveis tiveram seu escore Z próximo do escore máximo das variáveis-ruído introduzidas.

Figura 7 – Resultado da execução da seleção de variáveis



Fonte: Elaborado pelo autor

A decisão sobre a confirmação ou rejeição foi realizada através de um teste de diferença de médias em relação a estas variáveis-ruído. Foi estabelecido um limite de 100 iterações para o algoritmo, com nível de significância de 1% para determinar a diferença das médias necessárias para tomada de decisão.

Após a execução foram confirmadas 105 das 131 variáveis independentes, o que pode ser considerado um número alto, levando-se em conta a característica esparsa das muitas variáveis binárias. Das 30 variáveis não confirmadas, 24 foram rejeitadas e 2 não foram classificadas. As variáveis não classificadas foram arbitrariamente rejeitadas neste projeto, pois não foi possível durante o limite de iterações diferenciá-las das variáveis-ruído em relação à importância. Todas as quatro variáveis com valores faltantes imputados foram confirmadas como importantes.

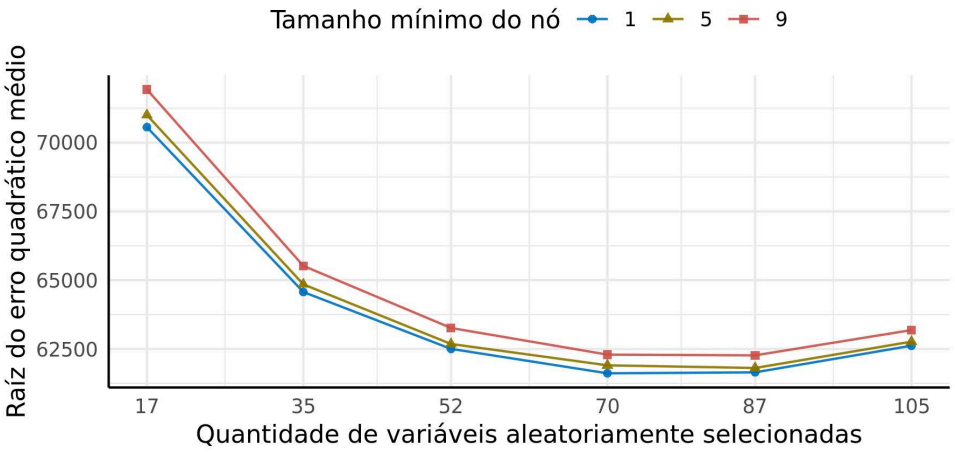
4.4 Modelo Final

Neste trabalho foi adotado a quantidade de 1000 árvores, sendo possível perceber melhor resultado em relação a 500 árvores, mas um aumento para 2000 árvores não provocou grandes diferenças nas métricas.

O gráfico da Figura 8 demonstra o RMSE para os modelos obtidos com as combinações

de número mínimo de nós e quantidade de variáveis aleatoriamente selecionadas. O modelo que obteve o menor RMSE possui tamanho mínimo do nó igual a 1 e 70 variáveis aleatoriamente selecionadas em cada particionamento. Estes valores foram utilizados no modelo final.

Figura 8 – Seleção de hiperparâmetros



Fonte: Elaborado pelo autor

As diferenças de resultados entre os modelos variando apenas o tamanho mínimo do nó não foram grandes. Os modelos com valor 1 obtiveram os melhores resultados, porém os resultados para o valor 5 não ficaram distantes. Utilizar um valor maior pode ser uma solução caso seja desejável reduzir a complexidade e tempo de construção das árvores geradas.

Para a troca da quantidade de variáveis aleatoriamente selecionadas o melhor valor foi obtido com 70, tendo o valor 87 resultado em um erro muito próximo. Uma maior diferença aparece apenas quando escolhemos valores menores que 52. A utilização de todas as 105 variáveis seria equivalente a um modelo *bagging* de árvores de regressão. Para este caso o resultado não foi o melhor, porém não foi tão distante do obtido na melhor combinação.

Através do processo computacional de busca em grade é possível destacar que os resultados com valores padrão já aparentam ser razoáveis, mas foi encontrada uma combinação de hiperparâmetros que gerou uma pequena melhoria. Este resultado está de acordo com Probst, Wright e Boulesteix (2019), que sugere que na maioria dos casos os valores definidos como padrão nas implementações de *Random Forest* já oferecem resultados satisfatórios.

4.5 Avaliação das Métricas

A avaliação do modelo foi realizada nos dados do conjunto de testes que inicialmente foi separado e somente agora utilizado, resultando nas métricas apresentadas na Tabela 5. Os resultados do modelo *Random Forest* foram comparados com os resultados de um modelo utilizando todas as variáveis em todas as etapas de particionamento, o que equivale a um simples *Bagging* de árvores de regressão.

Tabela 5 – Métricas do modelo.

Métrica	Random Forest	Bagging
RMSE	63.324,56	64.122,75
MAE	40.287,09	40.218,21
R^2	0,9392	0,9376
MAPE	8,21%	8,21%

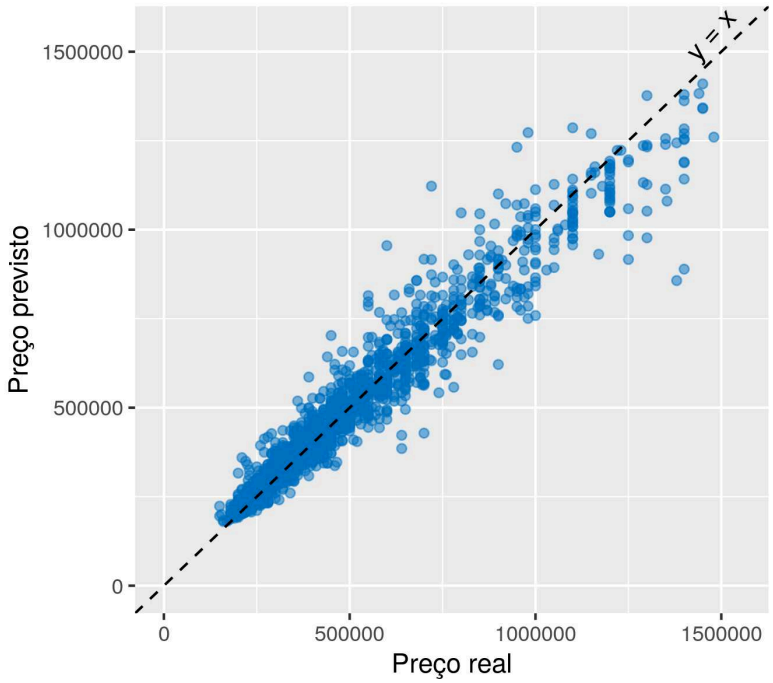
Para o modelo *Random Forest* treinado, o RMSE o valor obtido foi de R$ 63.324,56, este valor pode ser entendido como o desvio típico dos resíduos. O MAE obtido foi de R$ 40.287,09 e representa o erro médio absoluto. A unidade destas duas métricas é a mesma da variável dependente e sua escala depende da observação da distribuição destes valores para melhor perspectiva.

Uma análise mais intuitiva do resultado do modelo *Random Forest* pode ser feita através do R^2 e MAPE que indicam a qualidade do ajuste de forma percentual. Para o R^2 concluímos que o modelo explica 93,92% da variância da variável dependente. Através do cálculo da métrica MAPE foi calculado um valor de 8,21%, indicando que este é o erro percentual médio absoluto esperado para previsões futuras.

Comparando os resultados obtidos pelo modelo *Bagging* é possível perceber que para este conjunto de dados os valores não foram muito distantes, com vantagem para o modelo *Random Forest* que se saiu melhor em um maior número de métricas. Este resultado no conjunto de teste já era esperado, pois na Seção 4.4 foi possível ver o resultado da validação cruzada durante a seleção de hiperparâmetros.

Na Figura 9 verificamos a relação entre o preço dos imóveis no conjunto de teste e o preço previsto pelo modelo. É possível perceber que os resíduos têm maior amplitude para maiores preços, porém este comportamento não é impeditivo para o modelo *Random Forest*, pois homocedasticidade não é uma suposição.

Figura 9 – Valores reais vs. ajustados



Fonte: Elaborado pelo autor

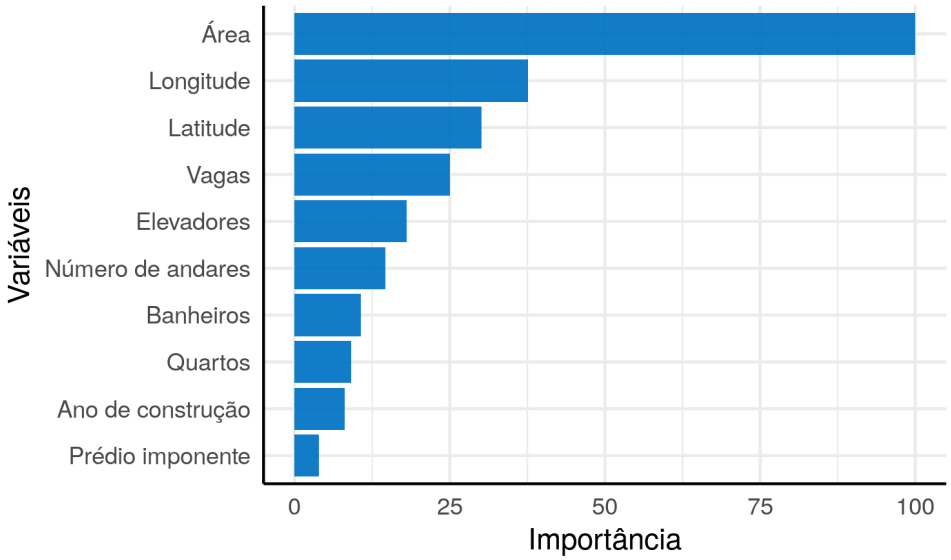
4.6 Importância das Variáveis

Estimativas das importâncias gerais das variáveis são obtidas através do método de permutação. Analisando a Figura 10, as 3 variáveis mais importantes são bastante intuitivas: área, latitude e longitude. Observando outras características do apartamento podemos destacar a importância do número de vagas de garagem e de banheiros. Analisando as características específicas do edifício é possível perceber que o número de andares, elevadores e o ano de construção tem forte influência sobre o preço. As duas variáveis imputadas que não foram apresentadas no gráfico ficaram dentre as 15 mais importantes.

Na Figura 11 observamos o resultado da avaliação das 10 variáveis mais importantes pelo método do impacto médio do SHAP. A ordem obtida foi quase idêntica ao do método de permutação, exceto pela troca de posição entre as variáveis latitude e vagas. Este método consiste de obter a média dos valores absolutos dos impactos de cada variável nas previsões. As variáveis classificadas como de maior importância consequentemente serão aquelas que geram a maior variação no preço do imóvel em relação ao que seria esperado na ausência desta variável no modelo.

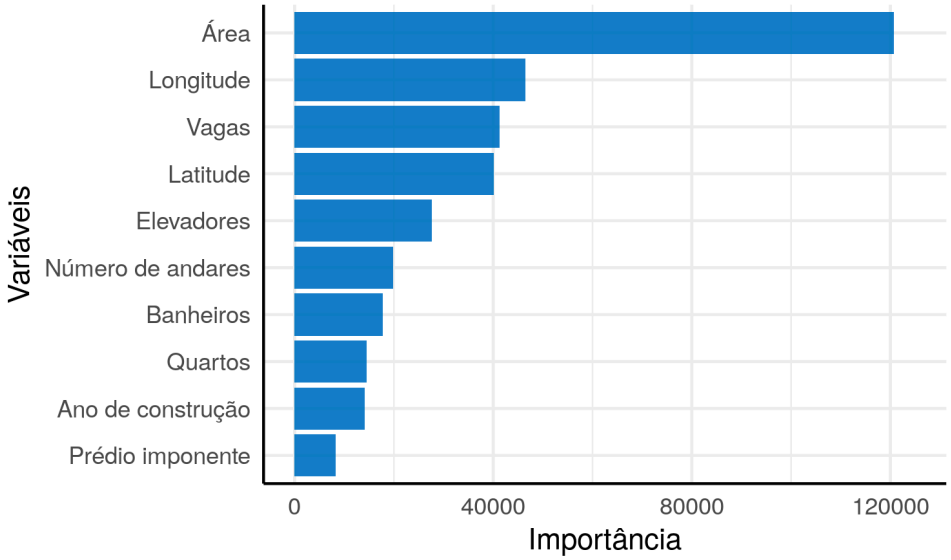
A importância das variáveis tem como objetivo apresentar uma visão geral das variáveis em relação ao quanto a remoção de cada uma degradaria o desempenho do modelo. Desta forma fica claro que apenas algumas variáveis têm uma grande importância geral para o modelo, porém as dezenas de variáveis restantes podem oferecer pequenas contribuições

Figura 10 – Importância das variáveis preditoras pelo método de permutação



Fonte: Elaborado pelo autor

Figura 11 – Importância das variáveis preditoras pelo método SHAP



Fonte: Elaborado pelo autor

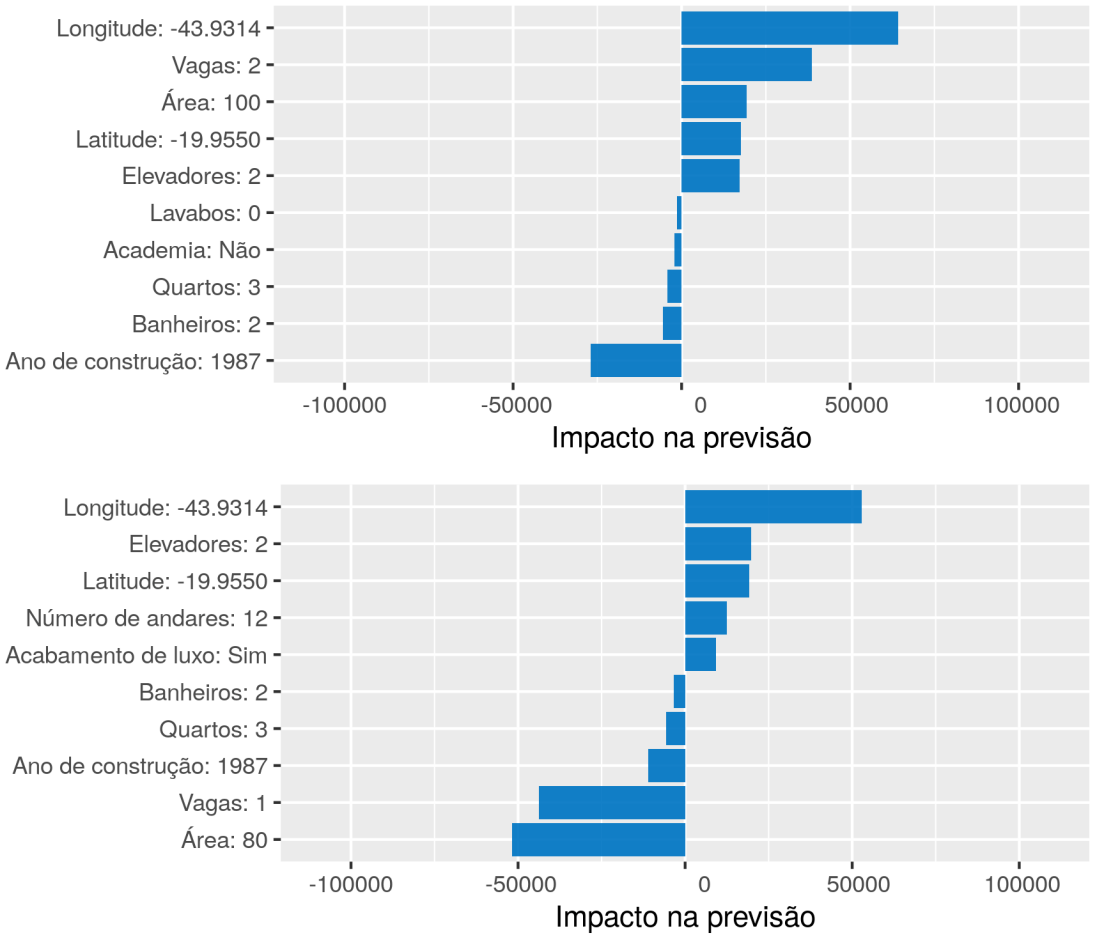
individuais para refinamento do resultado. Esta etapa de avaliação das importâncias não produziu controvérsias, pois não ocorreram divergências significativas entre os resultados nos dois métodos testados.

4.7 Explicação da Previsão

Uma previsão que pode ser explicada transfere maior confiança para quem deseja realizar qualquer ação que dependa do resultado de um modelo. Através do método *SHAP* podemos obter uma explicação de um previsão individual para um imóvel. Esta explicação justifica a amplitude e sentido da colaboração do valor presente em uma variável para o valor final previsto.

Na Figura 12 temos a explicação visual das variáveis de maior impacto, positivo e negativo, em duas previsões para um imóvel localizado no bairro Sion, com 3 quartos, 1 suíte e 2 banheiros. No primeiro gráfico o imóvel tem a área de 100 m^2 e 2 vagas de garagem, resultando em um preço previsto de R$ 686.166. No segundo gráfico o imóvel possui as mesmas características, exceto pela área reduzida para 80 m^2 e o número de vagas de garagem reduzido para 1, resultando em um preço previsto de R$ 524.034.

Figura 12 – Impacto na previsão de preço de dois imóveis localizados no mesmo endereço e com as mesmas características, exceto área e número de vagas (100 m^2 e 2 vagas para o de cima e 80 m^2 e 1 vaga para o de baixo)



Fonte: Elaborado pelo autor

Comparando as duas explicações é possível observar o comportamento do impacto na previsão das variáveis que foram alteradas. No primeiro gráfico as variáveis área e número de vagas de garagem impulsionam o modelo para uma previsão de maior valor. No segundo gráfico, já com os valores reduzidos, a situação se inverte e a área e número de vagas de garagem passam a ser as variáveis que mais impulsionam o modelo para uma previsão de menor valor. Portanto este formato de explicação permite repartir uma previsão nas contribuições individuais dos valores das variáveis.

Outro elemento para complementar a previsão são os intervalos de confiança para a estimativa média do preço previsto para o imóvel. Esta estimativa pode dar maior informação sobre o nível de incerteza do modelo do que uma única estimativa pontual. Para o modelo *Random Forest* estes intervalos podem ser obtidos através dos percentis das previsões individuais das árvores de regressão antes da agregação do resultado. Para o imóvel do primeiro gráfico, com preço previsto de R$ 686.166, podemos apresentar R$ 619.500 a R$ 780.000 como o intervalo de confiança de 90%. No segundo gráfico, com preço previsto de R$ 524.034, podemos apresentar o intervalo de R$ 409.995 a R$ 660.000.

5 Conclusão

Este trabalho abordou o problema da avaliação de preços de apartamentos em Belo Horizonte através da construção de um modelo preditivo com *Random Forest*. Utilizando também o método *SHAP* é possível entregar uma explicação da composição do preço previsto, aumentando a transparência do resultado.

A avaliação das métricas no conjunto de teste resultou em um erro médio absoluto percentual de 8,21% e um R^2 de 93,92%. Este resultado pode ser considerado satisfatório para o objetivo de apresentar uma avaliação imediata que sirva para que o proprietário tome imediatamente sua decisão sobre o preço pedido pelo apartamento.

O método de construção do modelo com *Random Forest* se mostrou bastante simples, pois nenhuma pressuposição precisa ser verificada. Outros pontos que tornam o método interessante é o pequeno número de hiperparâmetros para controlar o treinamento, a maior eficácia para lidar com grande número de variáveis e uma boa capacidade de generalização dos resultados. Em contrapartida, a incapacidade do método para lidar com dados faltantes é um ponto prejudicial, pois adiciona incertezas durante a etapa extra de imputação.

5.1 Trabalhos Futuros

Trabalhos futuros podem experimentar uma variação deste trabalho com transformação logarítmica da variável dependente, porém Feng et al. (2014) ressalta que esta transformação deve ser realizada com cuidado. Suas limitações devem ser conhecidas por parte do pesquisador, principalmente ao realizar análises em relação ao conjunto original dos dados.

Outra variação também pode ser realizada experimentando o método *Gradient Boosting*. Hastie, Tibshirani e Friedman (2009) apresenta um breve comparativo destes métodos em um conjunto de dados de imóveis e verifica que o *Gradient Boosting* melhora o desempenho com o aumento do número de árvores por um maior período.

Como complemento deste trabalho seria adequado também abordar intervalos de previsão com *RandomForest*, como apresentado em Meinshausen (2006), tendo a limitação de que algumas implementações não são escaláveis para grandes conjuntos de dados.

Referências

BREIMAN, L. Bagging predictors. **Machine Learning**, Springer Science and Business Media LLC, v. 24, n. 2, p. 123–140, 1996. Disponível em: <https://doi.org/10.1023/a:1018054314350>. Citado 2 vezes nas páginas 4 e 5.

BREIMAN, L. Random forests. **Machine Learning**, Springer Science and Business Media LLC, v. 45, n. 1, p. 5–32, 2001. Disponível em: <https://doi.org/10.1023/a:1010933404324>. Citado na página 5.

BREIMAN, L. et al. **Classification And Regression Trees**. Routledge, 1984. Disponível em: <https://doi.org/10.1201/9781315139470>. Citado 2 vezes nas páginas 3 e 4.

DONDERS, A. R. T. et al. Review: A gentle introduction to imputation of missing values. **Journal of Clinical Epidemiology**, Elsevier BV, v. 59, n. 10, p. 1087–1091, out. 2006. Disponível em: <https://doi.org/10.1016/j.jclinepi.2006.01.014>. Citado na página 8.

FENG, C. et al. Log-transformation and its implications for data analysis. **Shanghai archives of psychiatry**, v. 26, n. 2, p. 105—109, April 2014. ISSN 1002-0829. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4120293/>. Citado na página 28.

GENUER, R.; POGGI, J.-M.; TULEAU, C. Random forests: some methodological insights. v. 6729, 12 2008. Citado na página 7.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning**. Springer New York, 2009. Disponível em: <https://doi.org/10.1007/978-0-387-84858-7>. Citado 3 vezes nas páginas 4, 6 e 28.

HO, T. K. Random decision forests. In: **Proceedings of 3rd International Conference on Document Analysis and Recognition**. IEEE Comput. Soc. Press, 1995. v. 1, p. 278–282. Disponível em: <https://doi.org/10.1109/icdar.1995.598994>. Citado na página 5.

JAMES, G. et al. **An Introduction to Statistical Learning**. Springer New York, 2013. Disponível em: <https://doi.org/10.1007/978-1-4614-7138-7>. Citado 2 vezes nas páginas 3 e 5.

KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: **Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995. (IJCAI'95), p. 1137–1143. ISBN 1558603638. Citado na página 16.

KOHAVI, R.; JOHN, G. H. Wrappers for feature subset selection. **Artificial Intelligence**, Elsevier BV, v. 97, n. 1-2, p. 273–324, dez. 1997. Disponível em: <https://doi.org/10.1016/s0004-3702(97)00043-x>. Citado na página 9.

KUHN, M.; JOHNSON, K. **Applied Predictive Modeling**. Springer New York, 2013. Disponível em: <https://doi.org/10.1007/978-1-4614-6849-3>. Citado na página 3.

KURSA, M. B.; RUDNICKI, W. R. Feature selection with the boruta package. **Journal of Statistical Software**, Foundation for Open Access Statistic, v. 36, n. 11, 2010. Disponível em: <https://doi.org/10.18637/jss.v036.i11>. Citado na página 10.

KURSA, M. B.; RUDNICKI, W. R. The all relevant feature selection using random forest. **CoRR**, abs/1106.5112, 2011. Disponível em: <http://arxiv.org/abs/1106.5112>. Citado na página 10.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. In: GUYON, I. et al. (Ed.). **Advances in Neural Information Processing Systems 30**. Curran Associates, Inc., 2017. p. 4765–4774. Disponível em: <http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf>. Citado 2 vezes nas páginas 11 e 12.

MEINSHAUSEN, N. Quantile regression forests. **Journal of Machine Learning Research**, v. 7, n. 35, p. 983–999, 2006. Disponível em: <http://jmlr.org/papers/v7/meinshausen06a.html>. Citado na página 28.

NILSSON, R. et al. Consistent feature selection for pattern recognition in polynomial time. **J. Mach. Learn. Res.**, JMLR.org, v. 8, p. 589–612, maio 2007. ISSN 1532-4435. Citado na página 9.

PROBST, P.; BOULESTEIX, A.-L. To tune or not to tune the number of trees in random forest. **J. Mach. Learn. Res.**, JMLR.org, v. 18, n. 1, p. 6673–6690, jan. 2017. ISSN 1532-4435. Citado na página 7.

PROBST, P.; WRIGHT, M. N.; BOULESTEIX, A.-L. Hyperparameters and tuning strategies for random forest. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, Wiley, v. 9, n. 3, jan. 2019. Disponível em: <https://doi.org/10.1002/widm.1301>. Citado na página 22.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. Why should I trust you? In: **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. ACM, 2016. Disponível em: <https://doi.org/10.1145/2939672.2939778>. Citado na página 11.

STEKHOVEN, D. J.; BUHLMANN, P. MissForest–non-parametric missing value imputation for mixed-type data. **Bioinformatics**, Oxford University Press (OUP), v. 28, n. 1, p. 112–118, out. 2011. Disponível em: <https://doi.org/10.1093/bioinformatics/btr597>. Citado 2 vezes nas páginas 8 e 9.

WRIGHT, M. N.; KöNIG, I. R. Splitting on categorical predictors in random forests. **PeerJ**, PeerJ, v. 7, p. e6339, fev. 2019. Disponível em: <https://doi.org/10.7717/peerj.6339>. Citado na página 6.