Universidade de Brasília
Departamento de Estatística

Análise Imobiliária
Qual o melhor método para prever o valor de um imóvel?

Rafael Santana Araruna

Relatório final apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2022

Rafael Santana Araruna

Análise Imobiliária
Qual o melhor método para prever o valor de um imóvel?

Orientador(a): Leandro Tavares Correia

Relatório final apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2022

Agradecimentos

Primeiramente, gostaria de agradecer à minha família, que me proporcionou todos os recursos necessários, desde o início, para que eu tivesse uma educação de qualidade.

Gostaria de agradecer também aos meus amigos de curso, Amanda Shinkawa, Juliana Degani, Gabriel Peixoto, Ramon Moreira, Bruno Brandão e Matheus Erbisti, por sempre me ajudarem nas matérias quando eu estava com dificuldade e por sempre me darem apoio emocional pra seguir em frente e não desistir.

Gostaria de agradecer aos amigos de fora do curso, Maria Eduarda Ribeiro, Isadora Coelho, Juliana Eichler, Gabriel Coelho, Leonardo Gomes, Nicolau Ferraz, Pedro Torres, Ingrid Santos, Gabriela Magalhães e Ana Laura Pinheiro, por deixarem meus dias mais leves diante de todo o estresse da rotina e por me darem o suporte necessário para que eu chegasse até aqui.

Por fim, gostaria de agradecer ao meu orientador, Leandro Tavares, e também à professora Juliana Betini, que me ajudaram, com bastante paciência e eficiência, a desenvolver esse relatório.

Resumo

Tendo em vista que o ramo imobiliário cresceu bastante nos últimos anos, um dos efeitos que esse crescimento trouxe foi a escassez da mão de obra qualificada na área de comercialização, ou seja, na área responsável pela avaliação de imóveis.

Tal problemática pode ser justificada pelo fato da concorrência, neste setor, ter aumentado recentemente, o que levou alguns corretores, por inexperiência ou até para tentarem conquistar clientes, a se importarem somente com a prospecção do imóvel e não com sua avaliação dentro dos parâmetros impostos pelo mercado imobiliário. Por consequência, esses corretores realizam avaliações acima do mercado, ou seja, captam o imóvel e convencem o proprietário de que o valor é justo para o bem, quando, na verdade, é acima do real. Ações como essa prejudicam tanto os proprietários quanto os próprios corretores de imóveis em relação às negociações e ao mercado.

Dessa forma, o intuito deste trabalho é propor uma solução para essas falhas encontradas no ramo imobiliário em relação à precificação dos imóveis. Nesse sentido, serão construídos modelos de previsão por meio das técnicas de regressão linear, árvores de regressão, florestas aleatórias e redes neurais. Em seguida, será feita a comparação desses modelos através de determinadas métricas, com o objetivo de encontrar a metodologia estatística mais adequada e precisa para se prever o valor de um imóvel, evitando que o preço fique muito acima ou abaixo do esperado.

Portanto, uma vez que a avaliação é feita de forma correta, criteriosa e consciente, as negociações ocorrem com mais tranquilidade, eficiência e menos questionamentos. Isso é fundamental para que a comercialização seja realizada com transparência e segurança, rendendo bons frutos também para as empresas imobiliárias.

Palavras-chaves: regressão, modelo, árvore, floresta, rede, capacidade.

Lista de Tabelas

Lista de Figuras

Sumário

1 Introdução

Atualmente, percebe-se que as pessoas as quais desejam comprar ou vender imóveis possuem uma grande dúvida acerca de como estará o mercado imobiliário em um cenário pós-pandemia. É notório que a COVID-19 afetou vários setores da economia mundial e, consequentemente, o Brasil não conseguiu evitar esses prejuízos.

Em contrapartida, desde 2019, o ramo imobiliário apresenta sinais de evolução, indicando uma possível melhora nos anos subsequentes. De acordo com o site EXAME (2021), dados da Câmara Brasileira da Indústria da Construção (CBIC) apontaram para um aumento de 27,1% das vendas de imóveis no primeiro trimestre de 2021 em comparação com o mesmo período do ano anterior.

Além disso, segundo o site WIDESYS (2021), observa-se que, em 2021, o setor imobiliário foi um dos únicos que apresentou progresso, obtendo uma relevante alta em termos de vendas de imóveis quando em comparação com o ano de 2020. Esse setor, portanto, mostra que está se estabilizando, e as perspectivas são positivas para os próximos anos.

Diversos fatores explicam esse aquecimento do mercado imobiliário, como a alta disponibilidade de crédito, a baixa Taxa Selic, que implica em baixas taxas de juros, entre outros. Adicionalmente, outro motivo, o qual foi provocado pela pandemia e que pode ter sido o principal fator, trata-se da necessidade das pessoas terem de ficar em casa por longos períodos, em razão da alta contaminação e número de mortes. Dessa forma, em concordância com o site EXAME (2021), a pandemia provocou um impacto sobre a moradia e a relação com a residência. Devido a prática do isolamento social, além de ter que ficar mais tempo dentro de casa, muitas vezes trabalhando à distância, as pessoas começaram a considerar diferentes características, que, anteriormente, podiam ter menor relevância. Consequentemente, os indivíduos começaram a refletir sobre o quão confortável era a sua moradia e se deveriam buscar algo melhor e aconchegante.

Outrossim, seguindo o mesmo raciocínio, empresas dos mais diferentes ramos perceberam como a produtividade dos colaboradores é a mesma ou até maior em casa do que no local de trabalho, ao mesmo tempo em que elas diminuem os custos. Assim, de acordo com o site WIDESYS (2021), as pessoas, provavelmente, irão priorizar um lugar que seja exclusivamente designado ao *home office* no futuro, principalmente porque esse modo de trabalho deve perdurar após a pandemia.

Segundo o site EXAME (2021), a pesquisa Raio-X do FipeZAP emitiu que 46%

dos respondentes têm intenção de adquirir imóveis nos próximos três meses. Esse patamar está muito próximo do recorde já registrado (48%), e está bastante acima da média histórica da pesquisa (37%). Por isso, por enquanto, podemos vislumbrar um cenário favorável, e 2021 e 2022 devem ser anos interessantes para o desenvolvimento do mercado imobiliário.

Nesse sentido, tendo em vista que o ramo imobiliário cresceu bastante nos últimos anos, um dos efeitos que esse crescimento trouxe foi a escassez da mão de obra qualificada na área de comercialização, ou seja, na área responsável pela avaliação de imóveis. O processo de compra, venda ou de locação começa pela definição do valor de um determinado imóvel que será colocado no mercado. Assim, a importância da avaliação imobiliária passa a ser cada vez maior, pois as imobiliárias e os corretores deverão fazer, com responsabilidade, a avaliação do bem, visto que o correto valor fará com que o mercado possua a dinâmica e a liquidez esperada.

No entanto, pode-se dizer que o mercado imobiliário ainda possui deficiências em relação à avaliação de imóveis. Tendo em vista que a concorrência neste setor aumentou bastante, alguns corretores, por inexperiência ou até para tentarem conquistar clientes, se importam somente com a prospecção do imóvel e não com a sua avaliação dentro dos parâmetros impostos pelo mercado imobiliário. Por consequência, segundo o autor AMARY (2014), estes corretores realizam avaliações acima do mercado, captam o imóvel e convencem o proprietário de que o valor é justo para o bem, quando, na verdade, é acima do real. Além disso, outro fator que agrava ainda mais essa situação são os indíviduos os quais entram no mercado buscando resultados imediatos, ações essas que prejudicam tanto os proprietários quanto os próprios corretores de imóveis em relação às negociações e ao mercado.

Através da avaliação imobiliária é possível verificar se o imóvel está de acordo com o mercado, com o intuito de definir o seu valor justo para a comercialização. Afinal, quando um imóvel é avaliado corretamente, as chances de negociação do bem aumenta consideravelmente, evitando, assim, valores fora da realidade, fazendo com que o proprietário se prejudique por locar “barato demais” seu imóvel por exemplo, não conseguindo corrigir o preço depois do fechamento do contrato. Diversos proprietários desejariam ouvir que seu imóvel vale mais, contudo, trabalhar o imóvel com preços fora do mercado dificulta e, às vezes, até mesmo impede a negociação do mesmo.

Nesse sentido, a avaliação de imóveis é uma análise que leva em consideração tanto as propriedades do bem quanto os atributos externos e de mercado que contribuem para sua precificação justa e adequada. Dessa maneira, em conformidade com o site

BANIB (2019), uma vez que a avaliação é feita de forma correta, as negociações ocorrem com mais tranquilidade e eficiência, com menos questionamentos, isto é, uma avaliação criteriosa e consciente é fundamental para que o negócio seja realizado com transparência e segurança. Tal situação proporcionará bons frutos para as empresas imobiliárias, como:

- Definir o valor ideal do imóvel;
- Verificar quais são os pontos de melhorias para que ele seja vendido com mais facilidade;
- Identificar os pontos críticos, que podem determinar sua desvalorização futura;
- Garantir a segurança na negociação tanto para a imobiliária quanto para o proprietário e o comprador.

Portanto, o intuito desse trabalho é propor soluções para tais problemas encontrados no mercado imobiliário em relação à precificação dos imóveis, ou seja, o objetivo é encontrar a metodologia estatística mais adequada e precisa para se prever o valor de um imóvel, evitando que o preço fique muito acima ou abaixo do esperado. Simultaneamente, serão desenvolvidas técnicas específicas, como:

- A linguagem Python, através da coleta dos dados feita via *scraping* na plataforma Visual Studio Code;
- A linguagem R, através da implementação de técnicas estatísticas na plataforma RStudio;
- Técnicas de aprendizagem em máquina.

Para a realização do estudo serão utilizadas determinadas técnicas estatísticas de previsão, como:

- Modelo de Regressão Linear;
- Modelos não-paramétricos (Árvores de regressão, *Random Forest* e Redes Neurais).

É válido notar que as técnicas listadas acima foram utilizadas por várias pessoas em artigos e monografias, nos mais diversos temas. Nesse sentido, no intuito de entender melhor essas metodologias, saber como empregá-las e utilizá-las, o estudo em questão foi baseado em alguns trabalhos, por exemplo:

- Livro escrito por Izbicki e Santos (2020), que ensina bastante sobre Modelo de Regressão Linear, técnicas de validação e qualidade do ajuste, além de explicar diversos algorítmos computacionais e de mostrar como aplicá-los em estatística, como o Random Forest e as Redes Neurais;

- Artigo de PEREIRA J. C.; GARSON (2012), cujo objetivo é utilizar métodos de regressão linear múltipla na modelagem do preço de venda de casas da cidade de Sorocaba-SP com base em suas características;

- Trabalho de ROQUE (2009), o qual realiza um estudo sobre a empregabilidade da previsão do índice BOVESPA usando Redes Neurais Artificiais;

- Trabalho de Conclusão de Curso de SILVA (2021), que consiste em um estudo comparativo entre seis algoritmos de aprendizado de máquina, entre eles o Random Forest, sobre dados de reclamações, realizadas por meio do PROCON;

- Trabalho de COSSI (2017), que se trata do desenvolvimento de uma proposta para a previsão de ganho de peso em animais pelo método de regressão linear múltipla e uma técnica baseada em inteligência artificial, mais especificamente redes neurais artificiais.

2 Revisão de Literatura

2.1 Modelo de Regressão Linear Simples

De acordo com o site TRYBE (2022), regressão linear simples é uma espécie de modelo na estatística cujo objetivo é indicar qual será o comportamento de uma variável dependente (Y), ou também chamada de variável resposta, como uma função que contenha uma ou mais variáveis independentes (X), ou também chamadas de variáveis explicativas.

Nesse caso, utiliza-se apenas uma variável independente e uma dependente, com o intuito de verificar uma possível relação linear entre duas variáveis. Se tivermos mais que uma variável independente (X), utilizaremos a regressão linear múltipla.

Dessa forma, a fórmula de uma regressão linear simples é a seguinte:

$$Y = \beta_0 + \beta_1 X,$$

em que:

- Y corresponde à variável resposta, seria os valores previstos;
- X corresponde à variável independente;
- β_0 indica o valor de Y, quando X for igual a zero;
- β_1 corresponde à variável que determina o quão inclinada a reta será, pois ela determinará se a relação entre as variáveis é grande ou pequena.

2.2 Modelo de Regressão Linear Múltiplo

Uma generalização do modelo de regressão simples é o modelo de regressão múltiplo, no qual são consideradas mais de uma variável independente na equação. Dessa forma, a função será dada por:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i, \tag{2.2.1}$$

em que:

- Y_i é o valor da variável reposta para a i-ésima observação;

- $\beta_0, \beta_1, \ldots, \beta_p$ são os parâmetros desconhecidos;
- $X_{i1}, X_{i2}, \ldots, X_{ip}$ são constantes fixadas;
- ε_i são variáveis aleatórias independentes e com distribuição $N(0, \sigma^2), i = 1, \ldots, n$.

Os coeficientes são interpretados da mesma maneira como é feito na regressão linear simples: β_0 indica o valor esperado de Y_i se todas as variáveis X_{ip} ($i = 1, 2, \ldots, n$) forem nulas; β_j ($j = 1, 2, \ldots, p$) mostra a variação esperada de Y_i para um aumento de uma unidade na variável X_{ip} quando todas as outras variáveis explicativas são mantidas constantes; e ε_i corresponde ao erro aleatório associado à equação em estudo.

Além disso, a equação (2.2.1) pode ser definida pelos vetores de observação, pela matriz de variáveis explicativas, pelo vetor dos parâmetros e pelo vetor de erros:

$$Y_{n\times 1} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X_{n\times(p+1)} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}, \quad \beta_{(p+1)\times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \varepsilon_{n\times 1} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

Dessa forma, a equação (2.2.1) pode ser escrita em termos matriciais:

$$Y_{n\times 1} = X_{n\times(p+1)}\beta_{(p+1)\times 1} + \varepsilon_{n\times 1}. \tag{2.2.2}$$

2.2.1 Método dos Mínimos Quadrados

Este método, segundo Colaboradores da Wikipédia (2022), consiste em uma técnica de otimização matemática, e tem como objetico encontrar o melhor ajuste para um conjunto de dados, tentando, assim, minimizar a soma dos quadrados das diferenças entre o valor estimado e os dados observados (tais diferenças são chamadas resíduos).

Nesse sentindo, se trata de um estimador que minimiza a soma dos quadrados dos resíduos da regressão, de forma a maximizar o grau de ajuste do modelo aos dados observados.

2.2.2 Estimação dos parâmetros

Mantendo a notação matricial e utilizando os conhecimentos de estimação por mínimos quadrados, os coeficientes do modelo, ou seja, β_j, são estimados por:

$$\hat{\beta} = (X'X)^{-1}X'Y. \tag{2.2.3}$$

2.2.3 Valores Ajustados

Seja o vetor de valores ajustados $\hat{Y}_i$ denotado por $\hat{Y}$, então, em notação matricial, tem-se que:

$$\hat{Y} = X\hat{\beta}. \tag{2.2.4}$$

Nesse sentido, utilizando a definição da equação 2.2.3, a equação 2.2.4 pode ser reescrita como:

$$\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y.$$

2.2.4 Resíduos

Seja o vetor de resíduos $\hat{e}_i = Y_i - \hat{Y}_i$ denotado por $\hat{e}$, então, em notação matricial, tem-se que:

$$\hat{e} = Y - \hat{Y} = Y - X\hat{\beta}$$

ou,

$$\hat{e} = Y - \hat{Y} = Y - HY = (I - H)Y.$$

2.2.5 Estimador de σ^2

O estimador de σ^2 é dado por

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{e}_i^{\,2}}{n-(p+1)} = \frac{SQRes}{n-(p+1)} = MSRes,$$

em que SQRes significa Soma dos Quadrados dos Resíduos.

Porém, em termos matriciais, SQRes pode ser escrito como:

$$SQRes = Y'Y - \hat{\beta}'X'Y.$$

Assim, o estimador de σ^2, em termos matriciais, pode ser escrito como:

$$\hat{\sigma}^2 = \frac{Y'Y - \hat{\beta}'X'Y}{n - (p+1)}.$$

2.3 Testes de Hipóteses

Nessa seção, foi utilizado o livro BUSSAB Wilton de O. MORETTIN (2018) para obter as informações acerca das técnicas utilizadas.

2.3.1 Teste t

Esse teste consiste em verificar se o valor do parâmetro β_j é diferente de zero, isto é, testa se a variável X_j tem alguma influência sobre o valor esperado de Y_i. Para isso, estabelece-se as seguintes hipóteses:

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases}$$

O objetivo do teste é dizer se H_0 é ou não aceitável. Para isso, primeiramente, define-se o nível de significância α, o qual é definido da seguinte forma:

$$\alpha = P(rejeitar H_0 | H_0 verdadeira),$$

ou seja, é a probabilidade de rejeitar a hipótese nula dada que ela é verdadeira. Usualmente, o valor de α é fixado como 5%, 1% ou 0,1%.

Em seguida, calcula-se o p-valor, o qual indica a probabilidade de ocorrer valores da estatística mais extremos do que o observado, sob a hipótese de H_0 ser verdadeira. Assim, se o p-valor for menor que α, H_0 é rejeitada.

2.3.2 Teste de Correlação de Pearson

O teste de correlação de Pearson tem o intuito de verificar se existe associação entre as variáveis. As hipóteses são:

$$\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{cases}$$

em que ρ representa o coeficiente de correlação populacional.

A análise para verificar se H_0 é ou não aceitável é feita da mesma maneira como foi explicado no teste t.

Além disso, para quantificar a correlação entre as variáveis usa-se a seguinte medida:

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2)}\sqrt{(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2)}},$$

em que r corresponde ao coeficiente de correlação amostral.

2.3.3 Teste de Shapiro-Wilk

O objetivo desse teste é verificar a normalidade da variável resposta. Suas hipóteses são:

$$\begin{cases} H_0 : \text{A variável segue uma distribuição Normal} \\ H_1 : \text{A variável não segue uma distribuição Normal} \end{cases}$$

A análise para verificar se H_0 é ou não aceitável é feita da mesma maneira como foi explicado no teste t.

2.3.4 Teste de Durbin-Watson

Esse teste serve para verificar os desvios na direção de:

$$\varepsilon_i = p\varepsilon_{i-1} + v_i.$$

Explicando de uma maneira melhor, esse teste serve para detectar correlação de

ordem 1, auto-correlação de ordem 1, ou seja, seria verificar se o erro na unidade presente (ε_i) está correlacionado com o erro na unidade passada (ε_{i-1}).

De maneira resumida, seu objetivo é verificar se os erros são independentes ou não.

Suas hipóteses são:

$$\begin{cases} H_0 : \text{Os erros são independentes} \\ H_1 : \text{Os erros não são independentes} \end{cases}$$

A análise para verificar se H_0 é ou não aceitável é feita da mesma maneira como foi explicado no teste t.

2.3.5 Teste de Breusch-Pagam

Para verificar a homogeneidade da variância, pode-se aplicar o teste de Breusch-Pagan, que é utilizado para testar se a variância do erro de um modelo de regressão é constante. É indicado para grandes amostras e é sensível quanto à normalidade dos resíduos.

O teste possui as seguintes hipóteses:

$$\begin{cases} H_0 : \text{As variâncias dos erros são iguais} \\ H_1 : \text{As variâncias dos erros são diferentes} \end{cases}$$

A análise para verificar se H_0 é ou não aceitável é feita da mesma maneira como foi explicado no teste t.

2.4 Multicolinearidade

A multicolinearidade, de acordo com o site MINITAB (2019), consiste em um problema comum em regressões, no qual ocorre quando o modelo inclui variáveis explicativas correlacionadas não apenas com a sua variável de resposta, mas também entre elas.

Para essa metodologia usa-se o Fator de Inflação de Variância (VIF), o qual mede quanto a variância dos estimadores de mínimos quadrados é influenciada quando com comparada com variáveis explicativas que não são correlacionadas. Esse indicador é

definido da seguinte forma:

$$(VIF)_k = (1 - R_k^2)^{-1},$$

tal que R_k^2 é o coeficiente de determinação múltiplo da regressão de X_i com as demais $(p-1)$ variáveis explicativas.

Além disso, note que:

- Se $R_k^2 = 0 => (VIF)_k = 1$ e x_k não está correlacionada com as demais variáveis.
- Se $R_k^2 \neq 0 => (VIF)_k > 1$ e x_k está correlacionada com as demais variáveis.
- Se o máximo dos $(VIF)_k > 10$, implica que a multicolinearidade está influenciando as estimativas dos parâmetros.

Nesse sentido, também podemos fazer o cálculo da média dos $(VIF)_k$'s para tirar conclusões sobre a multicolinearidade. Esse cálculo é feito da seguinte forma:

$$(V\bar{I}F)_k = \frac{\sum_{k=1}^{p}(VIF)_k}{p},$$

tal que se $(V\bar{I}F)_k$ for um valor consideravelmente maior que 1, implica que a multicolinearidade está influenciando as estimativas dos parâmetros.

2.5 Observações Influentes

Uma observação é influente caso ela altere, de forma substancial, alguma propriedade do modelo ajustado (como as estimativas dos parâmetros, seus erros padrões, valores ajustados, etc). Nesse caso, usa-se uma medida de influêcia, chamada distância de Cook, a qual vai retorar um valor capaz de expressar o quanto uma particular observação afeta alguma propriedade do modelo.

2.5.1 Distância de Cook

Cook (1977,1979) sugeriu uma medida usando a distância ao quadrado entre todas as estimativas $\hat{\beta}$, e a estimativa obtida ao exluir a i-ésima observação, $\hat{\beta}_{(i)}$. Essa medida da distância é definida por

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})' M (\hat{\beta}_{(i)} - \hat{\beta})}{c},$$

em que $M = (X'X)$ e $c = (p+1)MSRes$. Logo:

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})' (X'X) (\hat{\beta}_{(i)} - \hat{\beta})}{(p+1)MSRes}.$$

A distância de Cook, medida da influência que a $i - sima$ observação tem sobre todos os n valores ajustados, também pode ser definida por:

$$D_i = \sum_{j=1}^{n} \frac{(\hat{Y}_j - \hat{Y}_{j_{(i)}})^2}{(p+1)MSRes}.$$

Além disso, a distância de Cook pode ser obtida usando apenas o resultado do ajuste do modelo com todos os dados por meio de

$$D_i = \frac{\hat{e}_i^2}{(p+1)MSRes} \left[\frac{h_{ii}}{(1-h_{ii})^2} \right],$$

onde $h_{ii} = X_i'(X'X)^{-1}X_i$. Assim, os valores de $D_i > 1$ são considerados observações influentes.

2.6 Seleção de Variáveis

Essa análise, a qual é feita sobre os dados de treinamento, consiste em verificar quantas variáveis o modelo deve possuir e, ao mesmo tempo, é verificado quais as melhores variáveis que devem entrar no modelo. Dessa forma, para fazer esse estudo, utiliza-se alguns critérios de seleção, como:

- O coeficiente de determinação múltiplo:

  $$R^2_{p+1} = 1 - \frac{SQRes_{p+1}}{SQT},$$

  em que SQT corresponde à Soma de Quadrados Total.

  Esse critério não penaliza o número de parâmetros, ou seja, ele sempre vai favorecer maior número de parâmetros no modelo, pois quanto maior o número de variáveis menor será o $SQRes_{p+1}$. Dessa foma, esse critério nos induz a adicionar mais variáveis. Além disso, para esse critério, quanto maior seu valor, melhor.

- O coeficiente de determinação múltiplo ajustado:

$$R^2_{a,p+1} = 1 - \frac{\frac{SQRes_{p+1}}{n-(p+1)}}{\frac{SQT}{n-1}} = 1 - \frac{MSRes_{p+1}}{MST}.$$

  Esse critério foi criado para resolver o problema do critério R^2_{p+1}, ou seja, ele penaliza modelos com um número grande de variáveis explicativas. Além disso, para esse critério, quanto maior seu valor, melhor.

- O critério C_{p+1} de Mallows's:

$$C_{p+1} = \frac{SQRes_{p+1}}{MSRes} - (n - 2p - 2).$$

  Esse critério penaliza modelos com um número grande de variáveis explicativas. Nesse caso, quanto menor o seu valor, melhor.

- O critério de Akaike - AIC_{p+1}:

$$AIC_{p+1} = nln(SQRes_{p+1}) - nln(n) + 2(p + 1).$$

  Esse critério, que surge como uma alternativa para os critérios anteriores, possui também penalidades para a adição de variáveis explicativas. Sua função de penalidade tem a finalidade de corrigir um viés proveniente da comparação de modelos com diferentes números de parâmetros. Nesse caso, quanto menor o seu valor, melhor.

- O critério de Informação Bayesiano - BIC_{p+1}:

$$BIC_{p+1} = nln(SQRes_{p+1}) - nln(n) + (p + 1)ln(n).$$

  Esse critério possui também penalidades para a adição de variáveis explicativas. Além disso, este penaliza mais modelos com maior número de parâmetros do que o critério AIC_{p+1}, tendendo, dessa forma, a selecionar modelos com um número menor de parâmetros. Nesse caso, quanto menor seu valor, melhor.

2.7 Análise da capacidade preditiva

Esse estudo, o qual é feito sobre os dados de validação, consiste em analisar os modelos de previsão quanto a sua capacidade de predição, ou seja, quanto ao seu ajuste

ou acurácia. Dessa maneira, para realizar esse estudo, utiliza-se, além do R^2 e do R_a^2 citados na seção anterior, as seguintes medidas:

- Critério $PRESS_{p+1}$:

$$PRESS_{p+1} = \sum_{i=1}^{n} (Y_i - \hat{Y}_{i,-i})^2.$$

  Esse critério, segundo Wikipedia contributors (2022), é uma forma de validação cruzada usada na análise de regressão para fornecer uma medida resumida do ajuste de um modelo a uma amostra de observações que não foram usadas para estimar o modelo. Nesse sentido, produzido um modelo ajustado, cada observação é removida e o modelo é reajustado usando as observações restantes. Assim, o valor previsto é calculado para a observação omitida em cada caso, e a estatística PRESS é calculada como a soma dos quadrados de todos os erros de previsão resultantes.

- Critério MSPR:

$$MSPR = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2.$$

  Esse critério consiste no erro quadrático médio de previsão, ou seja, se trata do risco preditivo. Apesar dele ser semelhante ao PRESS, seu cálculo é realizado sobre a base de validação e, além disso, não omite nenhuma observação, usa-se todas as observações da base.

2.8 Métodos Automáticos

2.8.1 Seleção Forward

Neste método, inicia-se com um modelo sem variáveis independentes, testando, passo a passo, a adição de uma nova variável com o uso de critérios de comparação para sua escolha (por exemplo, o teste t), adicionando a variável mais eficaz para o modelo, e repetindo este procedimento até não conseguir mais aumentar significativamente a acurácia do modelo.

2.8.2 Seleção Backward

Neste método, inicia-se a elaboração do modelo com todas as variáveis independentes, testando a eliminação de cada uma delas, usando um critério de comparação de escolha, eliminando as variávies que são menos eficazes para o modelo, e repetindo este procedimento até não ter mais melhoria no modelo.

2.8.3 Seleção Stepwise

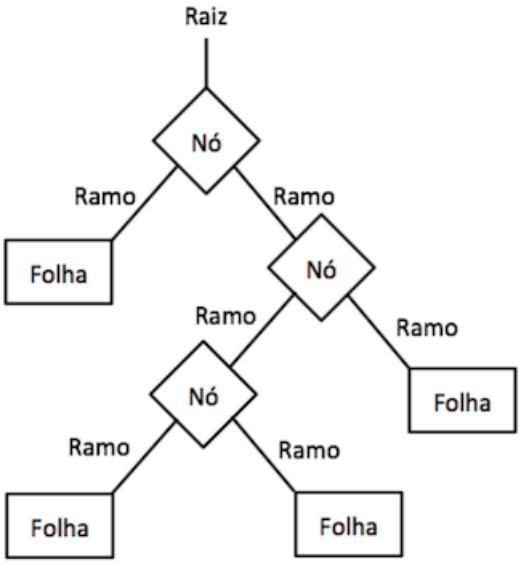
Consiste em um técnica de ajuste de modelos de regressão em que a escolha das variáveis preditivas é realizada por um procedimento automático. Em cada etapa, uma variável é considerada para adição ou subtração do conjunto de variáveis explicativas com base em algum critério pré-especificado. Normalmente, se assume a forma de uma sequência de testes t, mas outras técnicas são possíveis, como R_a^2, critério de informação de Akaike, critério de informação Bayesiano, Mallows, PRESS ou taxa de descoberta falsa. Resumindo, esse método aplica o Forward e o Backward ao mesmo tempo.

2.9 Árvores de Regressão

De acordo com o livro Izbicki e Santos (2020), as árvores de regressão consistem em uma forma de estimar a função de regressão, a qual vai fornecer um estimador bastante fácil de ser interpretado. Nessa metodologia, de maneira geral, a função de regressão estimada é constante por partes, isto é, a árvore particiona o espaço de covariáveis em pedaços menores.

A respeito da sua estrutura, uma árvore de regressão é dividida em quatro termos, conforme mostra a Figura 1 abaixo:

Figura 1: Ilustração de uma árvore de decisão



Fonte: SAKURAI (2018)

- Raíz: representa o nó inicial, o nó que fica no topo da árvore;
- Nós: cada nó interno da árvore corresponde a um teste do valor de uma propriedade;
- Ramos: os ramos dos nós são rotulados com os resultados possíveis do teste;
- Folha: cada folha da árvore especifica o valor a ser retornado se aquela folha for alcançada.

Acerca das suas vantagens, tem-se que:

- É uma estrutura muito fácil de ser aplicada e interpretada;
- Automaticamente, uma árvore está considerando interações, enquanto que na regressão linear, para incluí-las, é necessário colocar, explicitamente, o produto de uma variável com a outra no modelo;
- É muito fácil colocar variáveis categóricas em uma árvore, enquanto que em outras metodologias, outros estimadores, é preciso mais sofisticação;
- As árvores realizam seleção de variáveis automaticamente, ou seja, quando decidimos quais variáveis vão entrar em cada um dos nós, estamos deixando de fora, automaticamente, várias delas.

É importante ressaltar que a escolha de quais variáveis vão entrar nos nós está relacionada com o poder preditivo, ou seja, uma determinada variável é selecionada para entrar modelo quando esta consegue dividir bem os dados, isto é, criar folhas puras/homogêneas. Então, de certa forma, esta sendo aplicado uma seleção de variáveis.

Porém, uma desvantagem relevante dessa metodologia se trata das árvores serem muito simples para se ter um poder preditivo alto. Nesse sentido, com o intuito de obter um poder preditivo melhor, divide-se o espaço de covariáveis em uma partição $R_1, R_2, \cdots, R_j$. Uma vez feita essa partição, o próximo passo é realizar a predição em cada uma dessas partições. Essa predição consiste em basicamente ver todas as observações que se encaixam em uma determinada partição e, em seguida, calcular a média dos valores dessas observações.

No entanto, o verdadeiro problema está em como definir essas partições. Dessa forma, executa-se as seguintes etapas:

Etapa 1: Cria-se uma ávore "grande", ou seja, com muitas divisões/partições. As divisões da árvore são determinadas usando o critério/medida de pureza, que consiste no erro quadrático médio, isto é, na variância de cada folha somada. Assim, a divisão que for mais pura/homogênea será escolhida para entrar no modelo. No entanto, para construir a árvore é necessário realizar divisões binárias recursivas. Assim, a ideia é aplicar essas divisões e escolher, através da medida de pureza, qual a melhor divisão, até construir uma árvore grande. Geralmente, visto que o termo grande é relativo, para de se fazer as divisões quando chega-se em um ramo com menos de cinco observações. Porém, o fato de se ter uma árvore grande é que, provavelmente, causará *overfitting*, problema este que pode ser resolvido podando a árvore, que é a próxima etapa.

Etapa 2: Poda-se a árvore com o objetivo de evitar *overfitting* (variância alta). Podar uma árvore consiste em retirar cada nó/ramo da árvore, um por vez. Dessa maneira, é possível analisar, conforme se tira os nós/ramos, o que acontece com o erro estimado no conjunto de validação.

2.10 *Random Forest*

Acerca do *Random Forest* (Floresta Aleatória), cujas informações foram retiradas do livro Izbicki e Santos (2020), essa metodologia consiste em combinar modelos, ou seja, é um método de aprendizado conjunto, aonde vários modelos são criados a partir do

mesmo conjunto de dados e depois combinados de forma inteligente, produzindo melhores resultados do que a previsão baseada em um único modelo. A ideia, portanto, é combinar centenas de árvores de regressão, obtidas a partir de amostras do banco de dados, para chegar a uma melhor previsão, a qual é feita a partir da média dos valores obtidos, do que uma única árvore poderia fazer sozinha.

Outrossim, segundo a seção anterior, tem-se que que as árvores não possuem um poder preditivo muito bom, devido ao fato de serem muito simples. Porém, as florestas aleatórias utilizam a metodologia das árvores de regressão com o intuito de obter um poder preditivo melhor. Nesse sentido, para alcançar esse objetivo, é preciso do seguinte resultado teórico, o qual consiste na "Combinação de Predições".

Combinação de Predições

Esse algoritmo mostra que a combinação de duas funções de predição pode gerar uma terceira função de predição com um poder preditivo melhor. Ou seja, considerando duas funções de predição para Y, $g_1(x)$ e $g_2(x)$, o objetivo é verificar se, com a combinação dessas duas funções, é possível obter uma terceira função de predição $g_3(x)$ ainda melhor que elas, isto é, que tenha um poder preditivo melhor.

Dessa forma, sob certas condições, é possível chegar nesse resultado. As condições são:

- (i) Se $g_1(x)$ e $g_2(x)$ são não correlacionados;
- (ii) Se $g_1(x)$ e $g_2(x)$ são não viesados, ou seja, se não possuem alguma tendência para seguir determinado caminho ou agir de certa maneira;
- (iii) Se $g_1(x)$ e $g_2(x)$ têm a mesma variância.

Satisfazendo essas condições, tem-se que:

$$R(g_3(x)) \leqslant R(g_i(x)), \tag{2.10.1}$$

aonde $g_3(x) = \frac{g_1(x)+g_2(x)}{2}$. Ou seja, $g_3(x)$ tem um risco menor do que o risco de $g_1(x)$ e $g_2(x)$ separadamente. Assim, pode-se dizer que é melhor combinar funções.

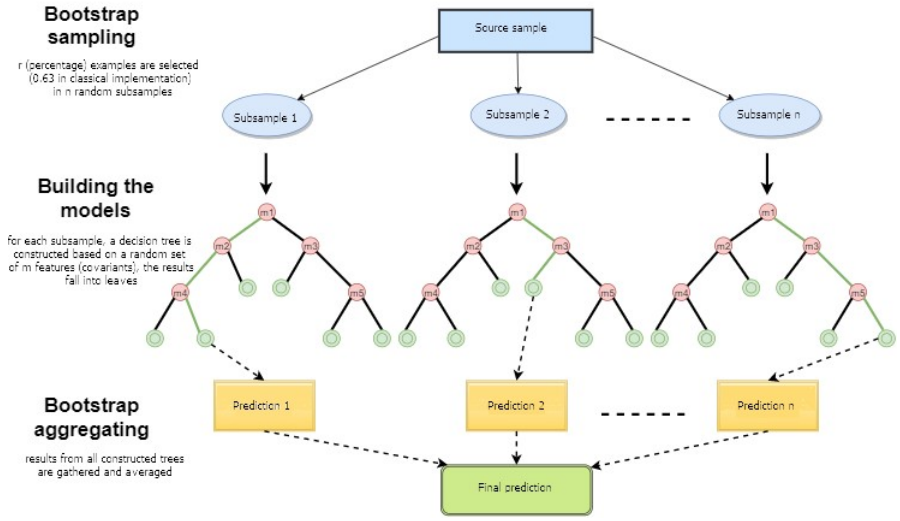
No entanto, as florestas aleatórias tentam, ainda que essas suposições não sejam satisfeitas, construir árvores de forma a se aproximar delas, ou seja, tentam criar $g_i(x)'s$ que satisfaçam essas suposições aproximadamente. Consequentemente, espera-se que (2.10.1) seja válido.

Portanto, de maneira resumida, a ideia do *random forest* é, basicamente, criar árvores de predição pro mesmo problema e combiná-las. No entanto, o processo de criação das árvores será diferente, quando comparado com o da seção anterior, para que as suposições (i) e (ii) sejam válidas aproximadamente.

Nesse sentido, em relação à criação das árvores, para que elas não sejam viesadas, é necessário não podá-las. Esta etapa de podar era feita para evitar o *overfitting*, ou seja, para diminuir a variância em troca do aumento do viés. Então, nesse caso, em que não se aplica o podamento das árvores, consequentemente tem-se uma variância grande, porém o viés será muito menor, e as árvores serão muito grandes também, isto é, profundas. Logo, consegue-se satisfazer, aproximadamente, a suposição (ii).

Além disso, acerca da suposição (i), se uma árvore ($g_1(x)$) não podada é construída para o conjunto de dados e, em seguida, uma segunda árvore ($g_2(x)$) também não podada é construída, chega-se exatamente na mesma árvore, ou seja, não tem nada de aleatório no processo de construção e, consequentemente, ambas as árvores serão totalmente correlacionadas ($g_1(x) = g_2(x)$). Assim, para diminuir essa correlação, as florestas aleatórias realizam o seguinte procedimento:

Figura 2: Esquema operacional da Floresta Aleatória



Fonte: MQL5 (2018)

Amostra Bootstrap

Ao invés de construir cada uma das árvores de regressão com a amostra original, as florestas aleatórias constroem cada uma delas com uma amostra bootstrap da amostra original, ou seja, uma amostra do mesmo tamanho do conjunto de dados, mas retirada

com reposição da amostra original. Nesse sentido, a ideia é criar B amostras bootstrap e, para cada uma delas, criar uma árvore não podada. Dessa forma, estrutura-se B árvores, cada uma criada com um conjunto de dados um pouco diferente. Consequentemente, tem-se funções de predição $g_i(x)'s$ diferentes entre si, ou seja, não correlacionadas, fato este que melhora bastante o desempenho das árvores e, assim, melhora seu poder preditivo.

Seleção de variáveis/Construção dos modelos

Para determinar qual variável irá entrar em cada divisão, não se observa todas as variáveis do banco de dados. Nesse caso, realiza-se um sorteio aleatório de m das d variáveis existentes em cada amostra bootstrap. Assim, será possível escolher, apenas entre essas m variáveis sorteadas, qual delas entrará na determinada divisão. Consequentemente, tem-se árvores ainda mais diferentes entre si, ou seja, menos correlacionadas ainda, pois as m variáveis sorteadas vão mudar, provavelmente, a cada sorteio. Nesse sentido, uma vez definido o m e, em seguida, a variável que vai entrar na primeira divisão, para a próxima divisão faz o mesmo processo e assim recursivamente, fazendo com que cada processo seja independente do outro.

Um observação importante de citar é que, após realizado o sorteio das m variáveis, a escolha da variável, dentre essas m sorteadas, que entrará em cada divisão é feita através da medida/critério de pureza, a qual foi mencionada anteriormente.

Então, de forma resumida, para cada nó/divisão, escolhe-se, através da medida de pureza, uma dentre m<d covariáveis, lembrando que cada subconjunto de covariáveis é escolhido aleatoriamente para cada nó/divisão.

2.11 Redes Neurais

Essa metodologia, segundo o livro Izbicki e Santos (2020), baseia-se também nas técnicas de árvores de decisão, as quais consistem em um mapa dos possíveis resultados de uma série de escolhas relacionadas, permitindo, assim, que um indivíduo ou organização compare possíveis ações com base em seus custos, probabilidades e benefícios. Além disso, pode ser usada tanto para conduzir diálogos informais quanto para mapear um algorítmo que prevê a melhor escolha, matematicamente.

As redes neurais, de maneira resumida, se trata de um conjunto de funções matemáticas (funções de ativação) que simulam o funcionamento do cérebro humano através da simulação de seus neurônios e de suas ligações. Nesse sentido, as redes neurais conecta cada função de ativação a várias outras de forma a ser capaz de transmitir uma informação

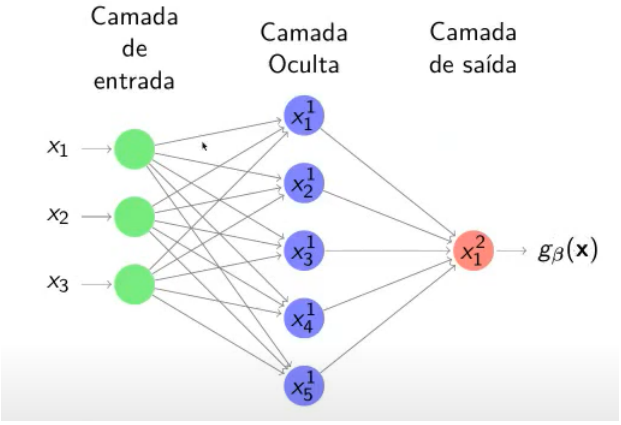
de maneira precisa após um processo de treinamento.

Desse modo, uma informação é passada ao sistema e, depois, é verificado se a mensagem interpretada pela rede está correta. Caso incorreto, os parâmetros são atualizados automaticamente e o processo é repetido até que todas as informações interpretadas estejam dentro de uma precisão esperada e aceitável. Além disso, uma vantagem das redes neurais é que sua arquitetura principal é sempre a mesma, mudando somente a quantidade de neurônios utilizados, em razão da quantidade de informação que se tem e da interpretação esperada. É válido falar também que seu processo de aprimoramento depende exclusivamente do conjunto de treinamento, ou seja, quanto maior o conjunto de treinamento, mais o sistema aprende.

Para entender melhor como funciona este processo, faz-se uma associação com a técnica de regressão linear. Nesse sentido, assim como esta especifica uma forma para a função de regressão, as redes neurais também especificam uma forma para essa função a qual se deseja estimar, porém, de uma forma bem mais complexa. Este procedimento se trata de uma família de funções, existindo a flexibilidade de escolher o quão complexa ela vai ser, pondendo ir desde uma regressão tão simples quanto uma regressão linear, até uma função extremamente complicada.

Além disso, a maneira como essa função é construída é interessante, pois permite que uma representação gráfica dessa função seja feita, conforme a Figura 3 mostra abaixo.

Figura 3: Estrutura de uma rede neural



Fonte: Imagem retirada da vídeo aula https://www.youtube.com/watch?v=b73pxvFvTV0

A Figura 3 está induzinho uma forma paramétrica para a função de regressão. Como foi dito anteriormente, tem-se uma flexibilidade muito grande de escolher essa estrutura, ou seja, há vários graus de liberdade que podem ser alterados, como aumentar

o número de camadas ocultas por exemplo, fato este que vai mudar a forma paramétrica que está sendo usada para aproximar a função de regressão. Nesse sentido, nota-se que essa classe é bastante complexa ao ponto de conseguir aproximar bem uma quantidade muito grande de funções. Devido a isso, considera-se a rede neural um procedimento não paramétrico.

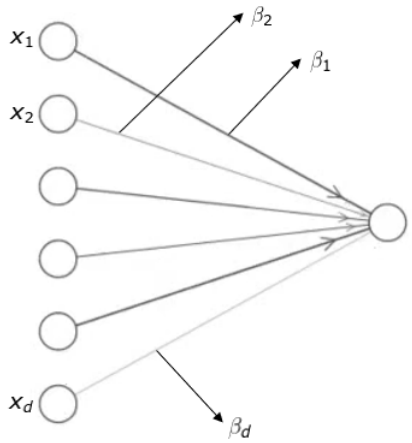
Outrossim, ainda acerca da Figura 3, tem-se que:

- Os neurônios verdes/camada de entrada representam covariáveis diferentes;
- Os neurônios roxos/camada oculta consistem em como compor as diferentes entradas para chegar na estimativa da função de regressão;
- Os neurônios vermelhos/camada de saída, representam uma função de x ($g_\beta(x)$).

É válido observar que é possível ter um ou mais pontos/neurônios vermelhos na camada de saída, e que nem toda estrutura terá a camada oculta.

Estrutura mais simples de Rede Neural

Figura 4: Estrutura de uma rede neural sem camada oculta



Fonte: Imagem retirada da vídeo aula https://www.youtube.com/watch?v=b73pxvFvTV0

A partir da Figura 4, a qual ilustra uma estrutura mais simples de rede neural, nota-se que esta possui apenas camada de entrada e saída. Esse tipo de estrutura consiste no caso mais simples de uma rede neural, pois não possui camada oculta, fato este que diminui sua complexidade e facilita sua interpretabilidade. Dessa forma, com o intuito de compreender mais sobre como funciona o mecanismo de uma rede neural, aproveita-se esse caso particular como primeiro exemplo, que é mais simples, e depois expande-se para situações mais complicadas.

Nesse sentido, acerca das flechas que saem dos neurônios da camada de entrada, sabe-se que:

- Para cada flecha tem-se um β_j associado;
- A informação que chega aos neurônios, que recebem essas flechas, é representada pela soma dos produtos entre o beta associado a cada flecha e o valor do neurônio anterior, como é mostrado abaixo:

$$\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_d x_d; \tag{2.11.1}$$

- Esses neurônios, que recebem as flechas, definem a informação que será retornada por eles aplicando uma função (função de ativação), a qual é pré-definida anteriormente, sobre a operação (2.11.1):

$$f(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_d x_d).$$

Assim, de maneira resumida, essa estrutura basicamente modela a função de regressão (que vai ser a saída da rede neural) como sendo uma função (f), que é escolhida de ante mão, aplicada nessa soma:

$$g(x) = f\left(\sum_{i=1}^{d} \beta_i x_i\right).$$

Imaginando o caso mais simples, em que a função de ativação pré-definida seja a função identidade, ou seja, $f(z) = z$, tem-se que a função de regressão é modelada através de:

$$g(x) = \sum_{i=1}^{d} \beta_i x_i.$$

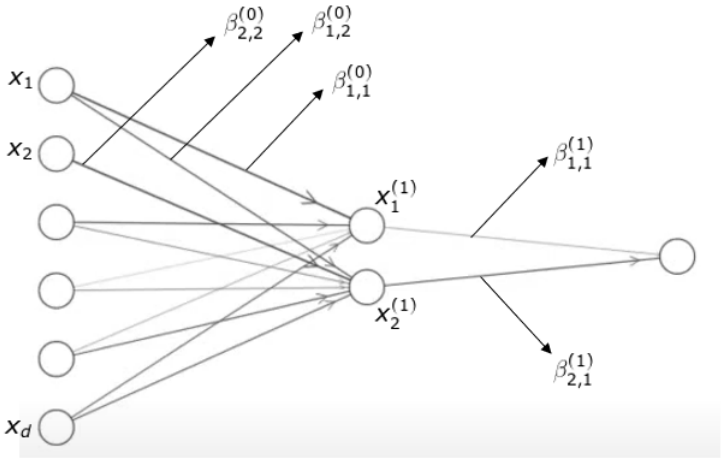
Nesse sentido, nota-se que se trata simplesmente de uma regressão linear, isto é, a rede neural, que está associada a essa estrutura, está basicamente impondo uma forma linear para a função de regressão.

Além disso, uma observação importante é que, uma vez que há um critério para estimar os betas, a forma da função de regressão estimada vai ser a forma de uma regressão linear ajustada. No entanto, não necessariamente esse critério consiste no método de mínimos quadrados, pois pode acontecer dos β's estimados não coincidirem com os betas

estimados por mínimos quadrados; mas ainda assim a estrutura da função de regressão ajustada vai ser linear.

Estrutura mais complexa de Rede Neural

Figura 5: Estrutura de uma rede neural com camada oculta



Fonte: Imagem retirada da vídeo aula https://www.youtube.com/watch?v=b73pxvFvTV0

Uma das vantagens da rede neural em relação à regressão linear é a possibilidade de tornar a estrutura mais complexa. Nesse sentido, como mostra a Figura 5, a qual ilustra uma rede neural mais complexa, nota-se que ao invés de ir direto da camada de entrada para a camada de saída, como foi visto no exemplo anterior, adiciona-se neurônios intermediários, ou seja, adiciona-se a camada oculta.

Acerca dos neurônios da camada oculta, sabemos que:

- Estes vão receber a combinação de todos os β's com os seus respectivos x_i's:

$$\sum_{i=1}^{d} \beta_{i,1}^{(0)} x_i,$$

em que $\beta_{i,1}^{(0)}$ representa todos os betas estimados, da primeira passagem, que vão para o neurônio intermediário $x_1^{(1)}$;

$$\sum_{i=1}^{d} \beta_{i,2}^{(0)} x_i,$$

em que $\beta_{i,2}^{(0)}$ representa todos os betas estimados, da primeira passagem, que vão para o neurônio intermediário $x_2^{(1)}$.

- Esses neurônios intermediários vão definir a informação que eles vão retornar aplicando a função de ativação (f), pré definida, sobre essa combinação, determinando, assim, novas covariáveis:

$$x_1^{(1)} = f\left(\sum_{i=1}^{d} \beta_{i,1}^{(0)} x_i\right);$$

$$x_2^{(1)} = f\left(\sum_{i=1}^{d} \beta_{i,2}^{(0)} x_i\right).$$

Nesse sentido, continuando o processo, dos neurônios intermediários $x_1^{(1)}$ e $x_2^{(1)}$, que representam as novas covariáveis, até o neurônio da camada de saída vão sair outras flechas, as quais representam novos betas/parâmetros estimados: $\beta_{1,1}^{(1)}$, que indica o beta da segunda passagem associado ao primeiro neurônio intermediário, o qual vai em direção ao neurônio de saída, e $\beta_{2,1}^{(1)}$, que indica o beta da segunda passagem associado ao segundo neurônio intermediário, o qual vai em direção ao neurônio de saída. Dessa maneira, tem-se que o neurônio da camada de saída vai receber essa combinação das novas covariáveis com os novos betas, ou seja, $\sum \beta_{i,1}^{(1)} x_i^{(1)}$. Assim, esse neurônio de saída vai estabelecer a informação que vai ser retornada aplicando a função de ativação (f), pré definida, sobre essa nova combinação, como é mostrado abaixo:

$$f\left(\sum_{i=1}^{d} \beta_{i,1}^{(1)} x_i^{(1)}\right).$$

Portanto, pode-se notar que essa estrutura de rede neural está parametrizando a função de regressão de uma maneira muito mais complexa que a anterior, visto que agora há uma composição de funções, funções estas que são não lineares.

Estimação dos β's

A estimação dos betas em redes neurais pode ser feita de maneira análoga ao que é feito na regressão linear, isto é, definir uma função objetivo e tentar minimizá-la, em relação aos β's, por meio da sua derivação. Porém, nesse caso, a função objetivo mais natural seria o EQM (Erro Quadrático Médio), ou seja, a soma dos erros ao quadrado, cuja fórmula é dada por

$$\sum_{i=1}^{n} (y_i - g_\beta(x_i))^2,$$

em que $g_\beta(.)$ representa a função do neurônio da camada de saída.

Dessa forma, tem-se como objetivo achar os β's ótimos, ou seja, que minimizam o somatório acima. Embasando-se na estrutura de rede neural mais simples, em que a função de ativação é a função identidade e sem camada oculta, é facil realizar a conta analiticamente, basta usar o estimador de mínimos quadrados. Porém, quando se trata de uma estrutura mais complexa, não se tem uma solução analítica, pois existe uma composição bastante complicada de funções. Nesse caso, para achar os β's ótimos, faz-se o uso do método numérico, aonde o método mais popular consiste no Back-Propagation.

O Back-Propagation, por sua vez, é basicamente um Gradiente Descendente, que é uma forma de otimizar funções. Este gradiente fornece valores iniciais para todos os betas, que são os pontos iniciais do algoritmo de otimização e, dado esses valores, em seguida, faz-se a atualização deles, como é mostrado abaixo:

$$\beta_{i,j}^{l} \leftarrow \beta_{i,j}^{l} - \eta \frac{\partial EQM(g_\beta)}{\partial \beta_{i,j}^{l}},$$

em que:

- $\beta_{i,j}^{l}$, que representa um parâmetro da rede da passagem "l" que liga o neurônio "i" da camada anterior ao neurônio "j", vai ser, em cada iteração, o valor anterior dele menos uma determinada constante η vezes o gradiente descendente, o qual deriva a função objetivo, que nesse caso é o EQM, em relação ao parâmetro $\beta_{i,j}^{l}$;

Supondo um exemplo para um caso muito particular, ou seja, considerando um β da última camada (H) e com a função de ativação sendo a função identidade ($f(z) = z$), tem-se o seguinte resultado:

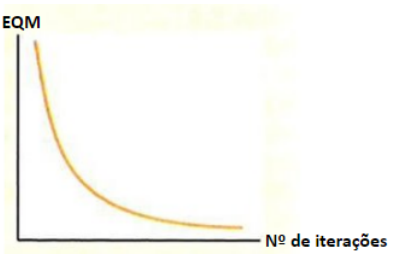
$$\frac{\partial EQM(g_\beta)}{\partial \beta_{j,1}^{H}} = \frac{\partial}{\partial \beta_{j,1}^{H}} \left[\frac{1}{n} \sum_{k=1}^{n} (\beta_{0,1}^{H} + \sum_{i=1}^{d_H} \beta_{i,1}^{H} x_{k,i}^{H} - y_k)^2 \right] = \frac{2}{n} \sum_{k=1}^{n} (\beta_{0,1}^{H} + \sum_{i=1}^{d_H} \beta_{i,1}^{H} x_{k,i}^{H} - y_k) x_{j,1}.$$

Assim, nota-se que a equação do gradiente, de forma geral, consiste em um formato de médias, ou seja, ele envolve a média das observações.

Portanto, de forma resumida, pode-se dizer que o gradiente descendente consiste em definir valores iniciais e, em seguida, fazer as iterações.

Outrossim, analisando a Figura 6, que ilustra a curva do EQM do número de iterações pelo valor do erro, é possível notar que:

Figura 6: Comportamento da curva do EQM conforme as iterações aumentam



- O erro diminui conforme o nº de iterações aumenta;
- Esse erro frequentemente vai chegar em zero ou muito próximo de zero, pois as redes neurais são estruturas super complexas, em particular, a dimensão do beta é maior que o tamanho amostral, ou seja, a rede neural tem muitos parâmetros;
- Assim, em um determinado momento os dados serão interpolados perfeitamente, isto é, o erro vai ser igual à zero; tal situação aparenta ser boa, pois a minimização foi alcançada, ou seja, chegou-se num ponto onde o erro é zero. No entanto, sabe-se que isso no geral não é bom, pois muitas vezes um estimador que interpola os dados, consequentemente, vai ter erro preditivo muito alto;
- dessa forma, costuma-se usar o que se chama de *earling-stop*, isto é, um método para parar mais cedo o aprendizado, então as iterações não são aplicadas até minimizar a função de fato;
- Portanto, separa-se uma parte do conjunto de dados para validação, e realiza-se, sobre ele, esse mesmo gráfico para acompanhar como o erro está variando; nota-se que, frequentemente, a curva do EQM começa alto, depois vai caindo até chegar em um certo ponto em que esta curva começa a crescer; esse ponto, aonde acontece essa mudança de direção, será o valor do *earling-stop*.

Em contrapartida, tem-se também o método do Gradiente Descendente Estocástico, o qual é usado com mais frequência atualmente e é, basicamente, uma variação do gradiente descendente. Acerca das nomenclaturas utilizadas neste método, tem-se:

- *Batch*: cada pedaço em que a amostra foi dividida;
- *Batch size*: tamanho de cada pedaço;
- *Epochs*: representa o número de passadas que se dá pelo treinamento inteiro ; quando se diz que quer treinar esse método usando cinco épocas, significa que irá ser fazer

o gradiente descendente estocástico e passar pelo banco inteiro cinco vezes, então cada observação vai entrar na conta de cinco gradientes.

Nesse caso, ao invés de usar a equação do gradiente que foi vista no método anterior, que é computacionalmente muito intensivo de ser calculado, a ideia é utilizar uma espécie de aproximação.

Nesse sentido, sabe-se que o gradiente, em casos muito gerais, para funções de objetivo muito gerais, tem esse formato de médias, ou seja, ele envolve a média de observações. Dessa maneira, visto que se tem uma média de observações, é possível pensar que, ao invés de usar todas as observações para fazer essa média, pode-se sortear, entre todas as observações, um subconjunto B. Assim, calcula-se o gradiente apenas para esse subconjunto B (*Batch*) sorteado, cuja fórmula é dada por

$$\frac{\partial EQM(g_\beta)}{\partial \beta_{j,1}^H} = \frac{2}{|\beta|} \sum_{k \in \beta} (\beta_{0,1}^H + \sum_{i=1}^{d_H} \beta_{i,1}^H x_{k,i}^H - y_k) x_{j,1}.$$

Tal realização torna o processo muito mais rápido e, além disso, consegue-se altenar entre quais *Batch*'s serão usadas em cada iteração. Dessa forma, esse é o gradiente descendente estocástico, e é estocástico porque existe aleatoriedade quando efetua-se o sorteio do subconjunto B. Portanto, de forma resumida, pode-se dizer que:

- Divide-se o banco de dados em partes;
- Em seguida, atualiza-se os parâmetros/gradiente usando a primeira parte, depois atualiza usando a segunda parte, depois usando a terceira, e assim por diante; ou seja, este procedimento mostra que há uma alternação no modo como é calculado o gradiente, pois não se calcula o gradiente com a amostra inteira, esta é utilizada em partes.

Então, esse método, além de ter vantagem computacional, tem boas propriedades teóricas. Aparentemente, redes treinadas com gradiente descendente estocástico fazem algum tipo de regularização na função a qual está sendo estimada, assim, quando ela é treinada com gradiente descendente estocástico, o problema de *overfitting*, que acontece com gradientes descendente normal, é muito menor. Logo, pode-se concluir que o gradiente descendente estocástico realmente melhora o desempenho, em termos de poder preditivo, da função estimada.

Observação

Existem várias outras melhorias para evitar/reduzir *overfitting*, como:

- *Dropout*: ao invés de calcular o gradiente usando todo mundo, basicamente considera-se que alguns neurônios não existem na hora de calculá-lo; é sorteado quais que não existem, assim, em cada passada realiza-se o sorteio de neurônios provavelmente diferentes, fato este que equivale a fazer a regularização;
- Regularização: ao invés de minimizar o EQM apenas, pode-se somar ao EQM uma penalização dos β's.

3 Metodologia

3.1 Material

Para a realização do estudo, primeiramente coletou-se um banco de dados via *scraping* do site DFIMÓVEIS, através da plataforma Visual Studio Code, cuja linguagem de programação usada é Python. Em seguida, executou-se uma limpeza, também em Python, do banco de dados obtido pelo *scraping*, ou seja, realizou-se o tratamento dos dados, que consiste em:

- Mudar e padronizar o nome das variáveis;
- Criar novas colunas (VALORM2);
- Retirar os valores faltantes (NA's);
- Retirar as observações repetidas;
- Extrair o tamanho da área, o valor, além do número de quartos, banheiros, suítes e vagas de cada imóvel;
- Extrair o estado, a cidade e o bairro de cada imóvel.

Após todos esses procedimentos, chegou-se no banco de dados final, o qual contém 359 anúncios de apartamentos à venda em Brasília, divididos em Asa Norte e Asa Sul, em que as variáveis são: ID, Descrição do Imóvel (DI), Tipo de Anúncio (TA), Tipo de Imóvel (TI), Localização do Imóvel (LI), ESTADO, CIDADE, BAIRRO, ÁREA, QUARTO, BANHEIRO, SUÍTE, VAGA, VALOR e Valor do m² (VALORM2). Segue, na Tabela 1, a descrição do conjunto de dados que será utilizado para as análises:

Tabela 1: Descrição do Banco de dados

Número	**Nome**	**Descrição**
1	ID	Número de identificação (1-150)
2	DI	Descrição do imóvel
3	TA	Tipo de Anúncio (Venda)
4	TI	Tipo de Imóvel (Apartamento)
5	LI	Localização do imóvel
6	ESTADO	Estado em que o imóvel se encontra
7	CIDADE	Cidade em que o imóvel se encontra
8	BAIRRO	Bairro em que o imóvel se encontra
9	ÁREA	Tamanho, em m^2, do imóvel
10	QUARTO	Número de quartos que o imóvel possui
11	BANHEIRO	Número de banheiros que o imóvel possui
12	SUÍTE	Número de suítes que o imóvel possui
13	VAGA	Número de vagas que o imóvel possui
14	VALOR	Valor, em R$, do imóvel
15	VALORM2	Valor do m^2 do imóvel (VALOR/ÁREA)

3.2 Método

Nesta seção, estrutura-se as etapas que serão aplicadas para analisar os dados descritos na Seção 3.1.

3.2.1 Análise Exploratória

Após a limpeza, faz-se, sob todo o banco de dados, uma análise descritiva dos dados, no RStudio, por meio de medidas descritivas (máximo, mínimo, média, mediana, variância, desvio padrão, coeficiente de variação, 1^o e 3^o quartil), de gráficos e tabelas, com o intuito de conhecer os dados que vão ser estudados, analisar como eles estão distribuídos e como estão se comportando.

Além disso, efetua-se também uma análise bidimensional, por meio de testes de correlação (Seção 2.3) e gráficos, visando verificar a existência de multicolinearidade.

3.2.2 Modelo Paramétrico

Nessa seção, será obtido um modelo paramétrico utilizando a técnica de Regressão Linear Múltipla (Seção 2.2).

Após obter uma primeira interpretação dos dados, o próximo passo é procurar um modelo que prevê, com precisão, o valor do imóvel em questão. Dessa maneira, primeiramente, divide-se o banco de dados em duas partes, dados treino e dados validação.

Acerca dos dados treino, efetua-se uma análise de regressão linear múltipla, também com o uso da ferramenta de programação RStudio. Nesta etapa, serão construídos alguns modelos:

- Modelo 1: composto por Área, Quarto, Banheiro, Súite, Vaga e Bairro, ambas na sua escala original, como sendo as variáveis explicativas, e Valor como sendo a variável resposta, também em sua escala original;
- Modelo 2: composto por Área, Quarto, Banheiro, Súite, Vaga e Bairro, ambas na sua escala original, como sendo as variáveis explicativas ; Valor como sendo a variável resposta, aplicada na escala logarítmica.

Vale ressaltar que, se na análise descritiva for constatado a possibilidade de classificar determinadas variáveis em categorias, outros modelos serão construídos com estas variáveis categorizadas.

Nesse sentido, sobre cada um desses modelos construídos, será realizada uma análise de regressão com o objetivo de obter as estimativas dos parâmetros (Seção 2.2), e verificar quais variáveis são significantes pelo teste t-Student (Seção 2.3), considerando um nível de significância de 5%. Dessa forma, após adquirir um modelo apenas com as variáveis relevantes, chamado de modelo final, é verificado se os modelos selecionados pelos critérios de seleção (Seção 2.6) e pelos métodos automáticos (Seção 2.8) convergem para este mesmo modelo final.

Ademais, por meio de gráficos e testes de hipóteses, efetua-se uma análise de diagnóstico, cujo propósito é examinar se este modelo final:

- Segue o pressuposto de Normalidade, utilizando o teste de Shapiro-Wilk (Seção 2.3);
- Segue o pressuposto de Independência dos erros, utilizando o teste de Durbin-Watson (Seção 2.3);

- Segue o pressuposto de Homocedasticidade, utilizando o teste de Breusch-Pagam (Seção 2.3);
- Possui valores influentes (Seção 2.5), utilizando a Distância de Cook;
- Apresenta Multicolinearidade (Seção 2.4).

Por fim, aplica-se a etapa de validação, que consiste em fazer uma análise de desempenho sobre dos modelos finais, cujo objetivo é examinar a qualidade do ajuste, ou seja, a capacidade preditiva de cada um deles. Dessa forma, utiliza-se as medidas da Seção 2.7, como o PRESS (Soma de Quadrados do Erro Residual Previsto), o MSPR (Erro Quadrático Médio de Previsão), o Coeficiente de determinação múltiplo (R^2) e o Coeficiente de determinação múltiplo ajustado (R_a^2), para comparar os modelos em questão. Em seguida, será definido o modelo de regressão final, que será aquele que possui a melhor capacidade de previsão, o melhor ajuste, ou seja, a melhor acurácia.

3.2.3 Modelos não Paramétricos

Nessa seção, o objetivo é tentar prever o valor do imóvel por meio de modelos não paramétricos, os quais serão construídos por meio das técnicas de: Árvores de regressão (Seção 2.9), *Random Forest* (Seção 2.10) e Redes Neurais (Seção 2.11).

3.2.4 Comparação dos Modelos de cada Metodologia

Obtendo-se o modelo paramétrico, que se trata do modelo de regressão linear, e os modelos não paramétricos, que se tratam dos modelos obtidos por Árvores de regressão, *Random Forest* e Redes Neurais, a próxima etapa seria compará-los em relação à precisão, capacidade preditiva e qualidade do ajuste. Assim, chega-se na metodologia mais adequada para se realizar a predição do preço de um imóvel. Dessa maneira, para fazer essa comparação, fez-se o uso do risco preditivo (MSPR), cujo cálculo encontra-se na Seção 2.7, de cada modelo construído.

4 Resultados

4.1 Análise Exploratória

Nessa seção, realiza-se a análise exploratória dos dados, com o objetivo de se ter uma primeira interpretação acerca do comportamento das variáveis que compõem o banco coletado. Dessa forma, efetua-se, por meio de gráficos e tabelas, tanto análises univariadas quanto bivariadas.

4.1.1 Análise da Variável Bairro

Nessa etapa, examina-se, através da Tabela 2, a qual fornece a frequência absoluta e relativa de cada bairro, a variável em questão.

Tabela 2: Tabela de Frequência da Variável Bairro

Bairro	Frequência Absoluta	Frequência Relativa
Asa Norte	214	59,61%
Asa Sul	145	40,39%
Total	359	100,00%

Analisando a Tabela 2, percebe-se que mais da metade das observações, 59,61%, estão localizadas no bairro Asa Norte, representando 214 apartamentos. Nesse sentido, os outros 40,39% se referem aos 145 apartamentos localizados no bairro Asa Sul.

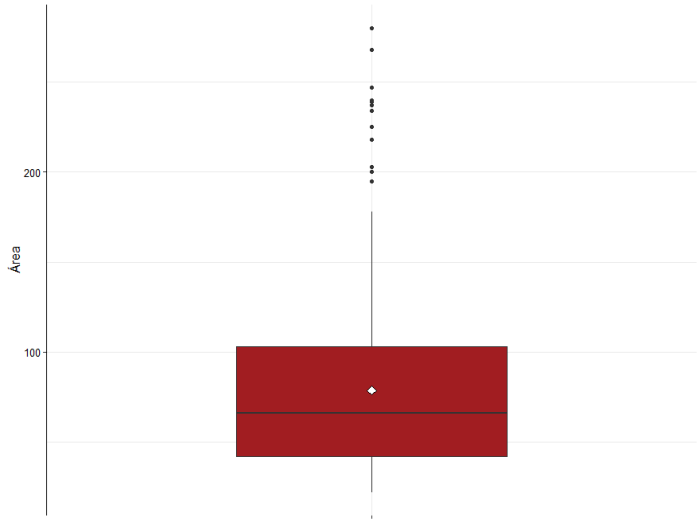
4.1.2 Análise da Variável Área

Nessa etapa, examina-se, através da Figura 7 e da Tabela 3, que mostram algumas medidas resumo da variável Área, a variável em questão.

Tabela 3: Medidas Resumo da Variável Área

Medidas	Mínimo	1º Quartil	Mediana	3º Quartil	Máximo	Média	Desvio Padrão
Valores	22	42	66	103	280	78,79	48,31

Figura 7: Box Plot da Variável Área



Analisando, simultaneamente, a Figura 7 e a Tabela 3, percebe-se que a área se distribui entre 22 m² (mínimo) e 280 m² (máximo), resultando em uma amplitude de 258 m². Além disso, nota-se uma média de 78,79 m² e um desvio padrão de 48,31 m², resultando em um coeficiente de variação de 61,21%, ou seja, alta dispersão dos dados em torno da média, indicando que as observações são heterogêneas (alta dispersão dos dados em torno da média). Observa-se, também, uma mediana de 66 m², isto é, metade dos dados está acima dela, e a outra metade está abaixo. Outrossim, é válido notar a presença de uma quantidade considerável de *outliers*, cujos valores são 280 m², 268 m², 247 m², 240 m², entre outros.

4.1.3 Análise da Variável Quarto

Nessa etapa, examina-se, através da Tabela 4, a qual fornece a frequência absoluta e relativa para cada quantidade de quartos, e da Tabela 5, a qual fornece algumas medidas resumo da variável Quarto, a variável em questão.

Tabela 4: Tabela de Frequência da Variável Quarto

Número de quartos	Frequência Absoluta	Frequência Relativa
1	141	39,28%
2	79	22,01%
3	109	30,36%
4	26	7,24%
5	4	1,11%
Total	359	100,00%

Analisando a Tabela 4, percebe-se que a maior parte dos apartamentos, 39,28% (141 apartamentos), possuem apenas 1 quarto, seguido dos apartamentos com 3 e 2 quartos, os quais representam, respectivamente, 30,36% (109 apartamentos) e 22,01% (79 apartamentos) de toda a amostra. Nesse sentido, pode-se dizer que há poucos apartamentos com 4 e 5 quartos, que representam, respectivamente, 7,24% (26 apartamentos) e 1,11% (4 apartamentos) de todo o conjunto de dados.

Tabela 5: Medidas Resumo da Variável Quarto

Medidas	Mínimo	1º Quartil	Mediana	3º Quartil	Máximo	Média	Desvio Padrão
Valores	1	1	2	3	5	2,09	1,04

Analisando a Tabela 5, nota-se uma média de 2,09 quartos e um desvio padrão de 1,04 quartos, resultando em um coeficiente de variação de 49,76%, ou seja, alta dispersão dos dados em torno da média, indicando que as observações são heterogêneas.

Outrossim, é válido notar que esta variável apresenta micronumerosidade, pois, como foi dito anteriormente, existem poucos imóveis com 4 e 5 quartos. Nesse sentido, sabendo que sua mediana é 2, seria viável dividir essa variável em duas categorias: possui 2 quartos ou menos, e possui mais de 2 quartos. Assim, tem-se dois grupos com uma boa quantidade de imóveis.

4.1.4 Análise da Variável Banheiro

Nessa etapa, examina-se, através da Tabela 6, a qual fornece a frequência absoluta e relativa para cada quantidade de banheiros, e da Tabela 7, a qual fornece algumas medidas resumo da variável Banheiro, a variável em questão.

Tabela 6: Tabela de Frequência da Variável Banheiro

Número de banheiros	Frequência Absoluta	Frequência Relativa
1	175	48,75%
2	97	27,02%
3	68	18,94%
4	7	1,95%
5	7	1,95%
6	4	1,11%
7	1	0,28%
Total	359	100,00%

Analisando a Tabela 6, percebe-se que a maior parte dos apartamentos, 48,75%

(175 apartamentos), possuem apenas 1 banheiro, seguido dos apartamentos com 2 e 3 banheiros, os quais representam, respectivamente, 27,02% (97 apartamentos) e 18,94% (68 apartamentos) de toda a amostra. Nesse sentido, pode-se dizer que há poucos apartamentos com mais de 3 banheiros, de modo que, se juntarmos todas essas observações, tem-se um total de apenas 19 apartamentos, ou seja, 5,29% de todo o conjunto de dados.

Tabela 7: Medidas Resumo da Variável Banheiro

Medidas	Mínimo	1º Quartil	Mediana	3º Quartil	Máximo	Média	Desvio Padrão
Valores	1	1	2	2	7	1,86	1,08

Analisando a Tabela 7, nota-se uma média de 1,86 banheiros e um desvio padrão de 1,078 banheiros, resultando em um coeficiente de variação de 57,95%, ou seja, alta dispersão dos dados em torno da média, indicando que as observações são heterogêneas.

Outrossim, é válido notar que esta variável apresenta micronumerosidade, pois, como foi dito anteriormente, existem poucos imóveis com mais de 4 banheiros. Nesse sentido, sabendo que sua mediana é 2, seria viável dividir essa variável em duas categorias: possui 2 banheiros ou menos, e possui mais de 2 banheiros. Assim, tem-se dois grupos com uma boa quantidade de imóveis.

4.1.5 Análise da Variável Suíte

Nessa etapa, examina-se, através da Tabela 8, a qual fornece a frequência absoluta e relativa para cada quantidade de suítes, e da Tabela 9, a qual fornece algumas medidas resumo da variável Suíte, a variável em questão.

Tabela 8: Tabela de Frequência da Variável Suíte

Número de suítes	Frequência Absoluta	Frequência Relativa
0	190	52,92%
1	150	41,78%
2	11	3,06%
3	4	1,11%
4	4	1,11%
Total	359	100,00%

Analisando a Tabela 8, percebe-se que mais da metade dos apartamentos, 52,92% (190 apartamentos), não possuem suíte, seguido dos apartamentos com apenas 1 suíte, que representam 41,78% (150 apartamentos) de toda a amostra. Nesse sentido, pode-se dizer que há poucos apertamentos com mais de 1 suíte, de modo que, se juntarmos todas

essas observações, tem-se um total de apenas 19 apartamentos, ou seja, 5,29% de todo o conjunto de dados.

Tabela 9: Medidas Resumo da Variável Suíte

Medidas	Mínimo	1º Quartil	Mediana	3º Quartil	Máximo	Média	Desvio Padrão
Valores	0	0	0	1	4	0,56	0,71

Outrossim, é válido notar que esta variável apresenta micronumerosidade, pois, como foi dito anteriormente, existem poucos imóveis com mais de 1 suíte. Nesse sentido, analisando a Tabela 9, nota-se que sua mediana é 0 e, assim, seria viável dividir essa variável em duas categorias: possui suíte, e não possui suíte. Dessa forma, tem-se dois grupos com uma boa quantidade de imóveis.

4.1.6 Análise da Variável Vaga

Nessa etapa, examina-se, através da Tabela 10, a qual fornece a frequência absoluta e relativa para cada quantidade de vagas, e da Tabela 11, a qual fornece algumas medidas resumo da variável Vaga, a variável em questão.

Tabela 10: Tabela de Frequência da Variável Vaga

Número de vagas	Frequência Absoluta	Frequência Relativa
0	205	57,1%
1	128	35,65%
2	19	5,29%
3	7	1,95%
Total	359	100,00%

Analisando a Tabela 10, percebe-se que mais da metade dos apartamentos, 57,1% (205 apartamentos), não possuem vaga, seguido dos apartamentos com apenas 1 vaga, que representam 35,65% (128 apartamentos) de toda a amostra. Nesse sentido, pode-se dizer que há poucos apertamentos com mais de 1 vaga, de modo que, se juntarmos todas essas observações, tem-se um total de apenas 26 apartamentos, ou seja, 7,24% de todo o conjunto de dados.

Tabela 11: Medidas Resumo da Variável Vaga

Medidas	Mínimo	1º Quartil	Mediana	3º Quartil	Máximo	Média	Desvio Padrão
Valores	0	0	0	1	3	0,52	0,69

Outrossim, é válido notar que esta variável apresenta micronumerosidade, pois, como foi dito anteriormente, existem poucos imóveis com mais de 1 vaga. Nesse sentido, analisando a Tabela 11, nota-se que sua mediana é 0 e, assim, seria viável dividir essa variável em duas categorias: possui vaga, e não possui vaga. Dessa forma, tem-se dois grupos com uma boa quantidade de imóveis.

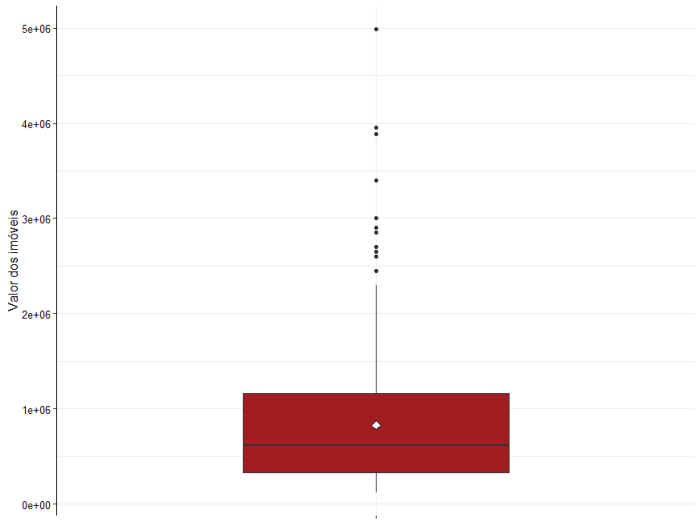
4.1.7 Análise da Variável Valor

Nessa etapa, examina-se, através da Figura 8 e da Tabela 12, que mostram algumas medidas resumo da variável Valor, a variável em questão.

Tabela 12: Medidas Resumo da Variável Valor

Medidas	Mínimo	1º Quartil	Mediana	3º Quartil	Máximo	Média	Desvio Padrão
Valores	120000	330000	620000	1160000	4990000	830141	664642,5

Figura 8: Box Plot da Variável Valor



Analisando, simultaneamente, a Figura 8 e a Tabela 12, percebe-se que o valor dos imóveis se distribui entre R$ 120000,00 (mínimo) e R$ 4990000,00 (máximo), resultando em uma amplitude de R$ 4870000,00. Além disso, nota-se uma média de R$ 830141,00 e um desvio padrão de R$ 664542,50, resultando em um coeficiente de variação de 80,06%, ou seja, alta dispersão dos dados em torno da média, indicando que as observações são heterogêneas. Observa-se, também, uma mediana de R$ 620000,00, isto é, metade dos valores está acima dela, e a outra metade está abaixo. Outrossim, é válido notar a presença de uma quantidade relevante de *outliers*, cujos valores são R$ 4990000,00, R$ 3950000,00,

R$ 3890000,00, R$ 3400000,00, entre outros.

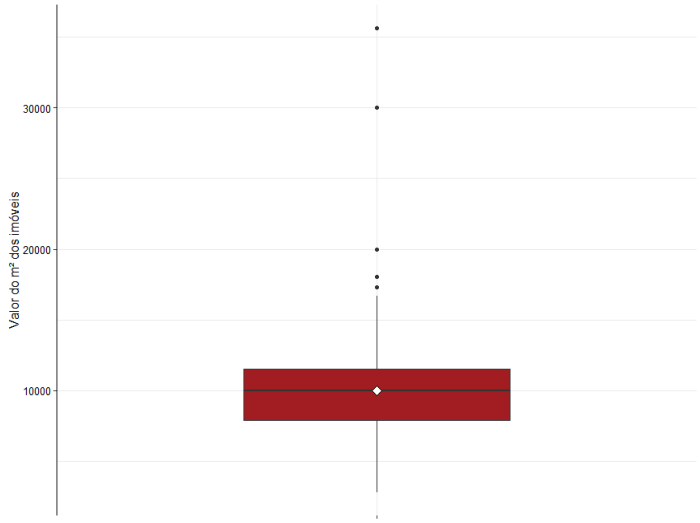
4.1.8 Análise da Variável Valor do m²

Nessa etapa, examina-se, através da Figura 9 e da Tabela 13, que mostram algumas medidas resumo da variável Valor do m², a variável em questão.

Tabela 13: Medidas Resumo da Variável Valor do m²

Medidas	Mínimo	1º Quartil	Mediana	3º Quartil	Máximo	Média	Desvio Padrão
Valores	2787	7910	10000	11538	35643	9997	3213,1

Figura 9: Box Plot da Variável Valor do m²



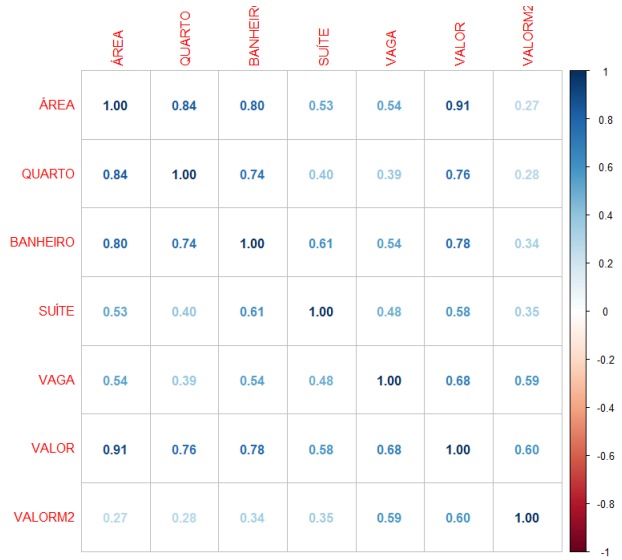
Analisando, simultaneamente, a Figura 9 e a Tabela 13, percebe-se que o valor do m² dos imóveis se distribui entre R$ 2787,00 (mínimo) e R$ 35643,00 (máximo), resultando em uma amplitude de R$ 32856,00. Além disso, nota-se uma média de R$ 9997,00 e um desvio padrão de R$ 3213,10, resultando em um coeficiente de variação de 32,14%, ou seja, entre média e alta dispersão dos dados em torno da média, indicando que as observações são um pouco heterogêneas. Observa-se, também, uma mediana de R$ 10000,00, isto é, metade dos valores está acima dela, e a outra metade está abaixo. Outrossim, é válido notar a presença de alguns *outliers*, cujos valores são R$ 35642,86, R$ 30000,00, R$ 20000,00, R$ 18051,11 e R$ 17333,33.

4.2 Análise Bidimensional

Nessa etapa, será feito uma análise conjunta de todas as variáveis explicativas quantitativas da amostra, no intuito de verificar quais variáveis estão relacionadas, e se essa relação é forte ou não, para, assim, dizer se há evidências de multicolinearidade.

Dessa forma, para se ter uma primeira interpretação dessa associação entre as variáveis em questão, faz-se a análise da seguinte matriz abaixo:

Figura 10: Matriz de Correlação



Analisando a Figura 10, percebe-se que há uma forte correlação positiva entre: Área e Quarto; Área e Banheiro; Área e Valor; Quarto e Banheiro; Quarto e Valor; Banheiro e Valor.

Dessa maneira, os resultados acima indicam a possível presença de multicolinearidade no conjunto de dados, fato este que será verificado mais adiante e, se for verdade, algumas variáveis irão ser desconsideradas para compor o modelo final.

4.3 Modelos Paramétricos

4.3.1 Modelo de Regressão 1

Nessa seção, será realizada a análise do modelo completo. Este, por sua vez, será formado por: Área, Quarto, Banheiro, Suíte, Vaga e Bairro, ambas na sua escala original, como sendo as variáveis explicativas ; Valor como sendo a variável resposta, também em sua escala original. Dessa forma, o modelo tem a seguinte estrutura:

$$Valor = \beta_0 + \beta_1 Area + \beta_2 Quarto + \beta_3 Banheiro + \beta_4 Suite + \beta_5 Vaga + \beta_6 Bairro_{(asasul)} + Erro$$

Por conseguinte, com o objetivo de verificar quais variáveis são significativas, aplica-se o teste t-Student sobre o modelo acima. Nesse sentido, todas as variáveis que possuírem um p-valor maior que o nível de significância, cujo valor é de 5%, serão retiradas do modelo. Assim, após retirar todas as variáveis que não foram significantes, chegou-se no seguinte modelo:

$$Valor = \beta_0 + \beta_1 Area + \beta_2 Suite + \beta_3 Vaga + Erro$$

Tabela 14: Análise do Modelo 1

Parâmetros	Estimativa	Erro Padrão	Estatística t	P-valor
β_0	-141932,4047	21661,0895	-6,55	0,0000
β_1	10504,3189	286,0464	36,72	0,0000
β_2	49058,3452	18863,4387	2,60	0,0100
β_3	208568,7237	20717,8885	10,07	0,0000

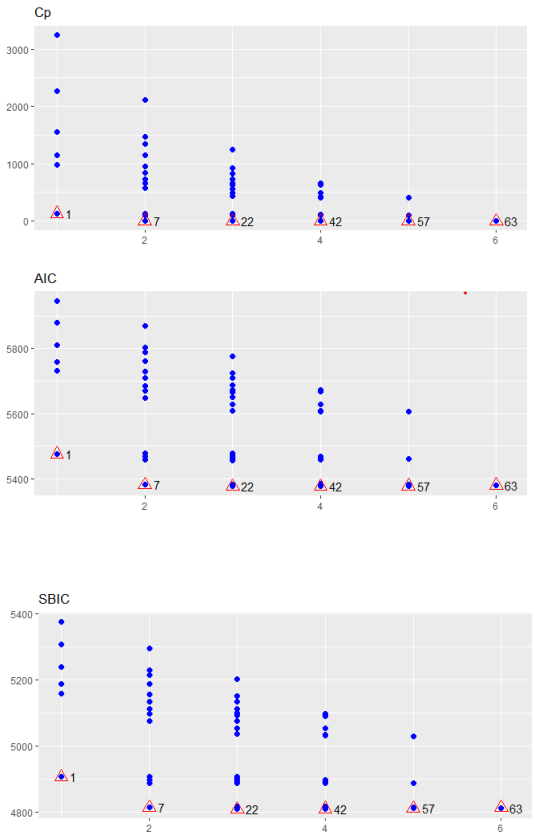
Observando a Tabela 14, nota-se que todas as variáveis são significantes, pois apresentaram um p-valor menor que o α, ou seja, foram retiradas as variáveis Quarto, Banheiro e Bairro. Dessa forma, tem-se o Modelo 1 final, e a próxima etapa é verificar se os modelos selecionados pelos critérios de seleção e pelos métodos automáticos convergem para este mesmo modelo.

Modelo selecionado pelos critérios de seleção

Nesta etapa, o objetivo é utilizar determinados critérios de seleção, como o critério de Mallows, de Akaike e de Informação Bayesiano, os quais penalizam a quantidade de variáveis, para verificar qual o melhor número de variáveis para entrar no modelo e, assim, verificar se converge para a mesma quantidade do Modelo 1 final definido anteriormente,

que possui 3 variáveis. Dessa forma, faz-se a análise da Figura 11, que mostra qual a melhor quantidade de variáveis, selecionada por cada critério de seleção, para compor o modelo.

Figura 11: Gráficos dos critérios de seleção



Observando a Figura 11, percebe-se que as melhores quantidades para compor o modelo são 2, 3, 4 e 5, ambas estão com valores próximos. Porém, as quantidades 3 e 4, apesar da pouca diferença, possuem as menores medidas e, assim, pode-se dizer que elas são as melhores escolhas.

Além disso, por meio de alguns critérios de seleção, também é possível saber qual o melhor modelo para cada quantidade de variáveis selecionada anteriormente. A partir disso, foi feito uma tabela com o *ranking* dos três melhores modelos com 3 e 4 variáveis, como é apresentado na Tabela 15.

Tabela 15: *Ranking* dos três melhores modelos com 3 e 4 variáveis

Ranking	Nº de variáveis	Bairro	Área	Quarto	Banheiro	Suíte	Vaga	R^2	R^2_a	C_p	BIC
1º	3	0	1	0	0	1	1	0,95	0,94	3,51	-560,72
2º	3	0	1	0	1	0	1	0,94	0,94	6,92	-557,26
3º	3	1	1	0	0	0	1	0,94	0,94	9,38	-554,81
1º	4	1	1	0	0	1	1	0,95	0,94	3,95	-557,03
2º	4	0	1	0	1	1	1	0,95	0,94	4,53	-556,43
3º	4	0	1	1	0	1	1	0,95	0,94	5,34	-555,59

A partir dos resultados da Tabela 15, observa-se que os melhores modelos com 3 e 4 variáveis, respectivamente, são:

$$Valor = \beta_0 + \beta_1 Area + \beta_2 Suite + \beta_3 Vaga + Erro$$

$$Valor = \beta_0 + \beta_1 Area + \beta_2 Suite + \beta_3 Vaga + \beta_4 Bairro_{(asasul)} + Erro$$

Dessa maneira, comparando os dois modelos acima em relação aos critérios C_p e BIC, percebe-se que o modelo com 3 variáveis possui as menores medidas, ou seja, ele é melhor que o modelo com 4 variáveis, convergindo, assim, para o mesmo Modelo 1 final.

Modelo selecionado pelos métodos automáticos

Nesta etapa, o objetivo é utilizar os métodos automáticos, que são o Forward, o Backward e o Stepwise, para verificar qual modelo eles selecionam como o melhor e, assim, verificar se este converge para o mesmo Modelo 1 final definido anteriormente, da mesma forma como o modelo de seleção de variáveis convergiu. Dessa maneira, com o uso da plataforma RStudio, foi constatado que ambos os métodos convergiram para o mesmo modelo, que foi:

$$Valor = \beta_0 + \beta_1 Area + \beta_2 Suite + \beta_3 Vaga + Erro$$

Nesse sentido, percebe-se que o modelo selecionado pelos métodos automáticos convergiu para o mesmo Modelo 1 final.

Portanto, pode-se dizer que, considerando um modelo sem transformação na variável resposta e com todas variáveis explicativas na sua escala original, o modelo mais adequado é o modelo composto por Área, Suíte e Vaga. No entanto, ainda é necessário realizar a análise de diagnóstico deste modelo para examinar se este atende a todos os pressupostos.

Análise de diagnóstico do Modelo 1 final

Primeiramente, realiza-se a verificação dos três pressupostos:

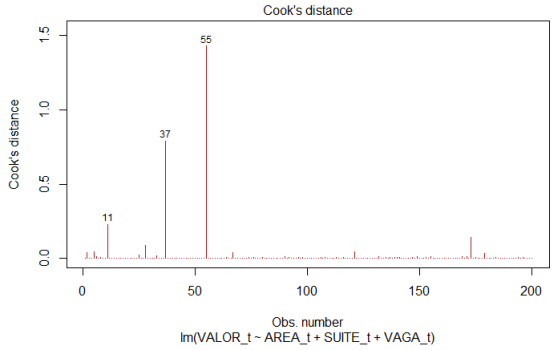
Tabela 16: Resultado dos Testes

Pressupostos	Testes	P-valor
Normalidade	Shapiro-Wilk	$< 0{,}001$
Independência dos erros	Durbin-Watson	0,4397
Homogeneidade da variância	Breusch-Pagan	$< 0{,}001$

A partir dos resultados da Tabela 16, nota-se que, considerando um nível de significância de 5%, os testes de normalidade e homogeneidade da variância resultaram em um p-valor menor que o α, ou seja, rejeitaram suas respectivas hipóteses nulas. Assim, há evidências para se dizer que os dados não possuem distribuição normal e nem variâncias iguais. Tal fato pode ser justificado pela presença de *outliers* no conjunto de dados e, assim, mais a frente será construído um novo modelo sem esses valores discrepantes e será verificado novamente estes pressupostos.

Outrossim, é importante examinar a presença de observações influentes, pois isto também pode ter prejudicado a validação dos pressupostos. Nesse sentido, utiliza-se da metodologia chamada DFCOOKS, a qual analisa a influência sobre o valor ajustado geral. Dessa forma, faz-se o estudo do gráfico abaixo:

Figura 12: Gráfico para verificar observações influentes



Analisando a Figura 12, é possível notar que existem três observações possivelmente influentes, que são: 11, 37 e 55. Além disso, é importante observar que há outras observações com essa possibilidade, mas que não foram destacadas no gráfico como as três citadas anteriormente, por exemplo: 2, 5, 25, 28, 67, 121, 173 e 179, totalizando, assim, em 11.

Ademais, é válido analisar também se existe multicolinearidade. Dessa forma, calcula-se os VIF_k's de cada variável do modelo, cujos resultados foram:

Tabela 17: Resultado dos VIF_k's

Variável	VIF_k's
Área	1,700080
Suíte	1,447153
Vaga	1,667451

A partir da análise da Tabela 17, nota-se que ambas as variáveis possuem VIF próximo de 1, ou seja, essas variáveis não estão correlacionadas entre si. Além disso, anali-

sando o VIF médio, cujo valor é de 1,60, percebe-se que não é um valor consideravelmente maior que 1 e, assim, pode-se dizer que a multicolinearidade não está influenciando as estimativas dos parâmetros.

Construção do modelo sem os valores discrepantes e influentes

Como foi visto na análise de diagnóstico, observou-se que este modelo rejeitou os pressupostos de normalidade e homogeneidade da variância, fato este provavelmente causado pela presença de valores discrepantes e influentes. Assim, decidiu-se tirar todas essas observações, para verificar se, sem esses dados, o modelo atende a todos os pressupostos. Dessa forma, realizando novamente os testes, obteve-se os seguintes resultados:

Tabela 18: Resultado dos Testes

Pressupostos	Testes	P-valor
Normalidade	Shapiro-Wilk	0,5223
Independência dos erros	Durbin-Watson	0,6297
Homogeneidade da variância	Breusch-Pagan	0,08323

A partir da análise da Tabela 18, percebe-se que todos os testes resultaram em um p-valor maior que o nível de significância α, cujo valor é de 5%. Assim, pode-se dizer que não há evidências suficientes para rejeitar as hipóteses nulas de cada teste, corroborando com a ideia de que esses dados estavam prejudicando o ajuste do modelo. Portanto, conclui-se que este modelo, após tirar todos os valores discrepantes e influentes, atendeu a todos os pressupostos e, assim, é correto dizer que ele é o mais adequado para essa estrutura.

4.3.2 Modelo de Regressão 2

Nessa seção, será realizada a análise do modelo completo, porém, será empregado o logarítmo na variável resposta. Nesse sentido, este modelo será formado por: Área, Quarto, Banheiro, Suíte, Vaga e Bairro, ambas na sua escala original, como sendo as variáveis explicativas ; Valor como sendo a variável resposta, aplicada na escala logarítmica. Dessa forma, o modelo tem a seguinte estrutura:

$$Log(Valor)=\beta_0+\beta_1 Area+\beta_2 Quarto+\beta_3 Banheiro+\beta_4 Suite+\beta_5 Vaga+\beta_6 Bairro_{(asasul)}+Erro$$

Por conseguinte, com o objetivo de verificar quais variáveis são significativas,

aplica-se o teste t-Student sobre o modelo acima. Nesse sentido, todas as variáveis que possuírem um p-valor maior que o nível de significância, cujo valor é de 5%, serão retiradas do modelo. Assim, após retirar todas as variáveis que não foram significantes, chegou-se no seguinte modelo:

$$Log(Valor)=\beta_0+\beta_1 Area+\beta_2 Quarto+\beta_3 Vaga+\beta_4 Bairro_{(asasul)}+Erro$$

Tabela 19: Análise do Modelo 2

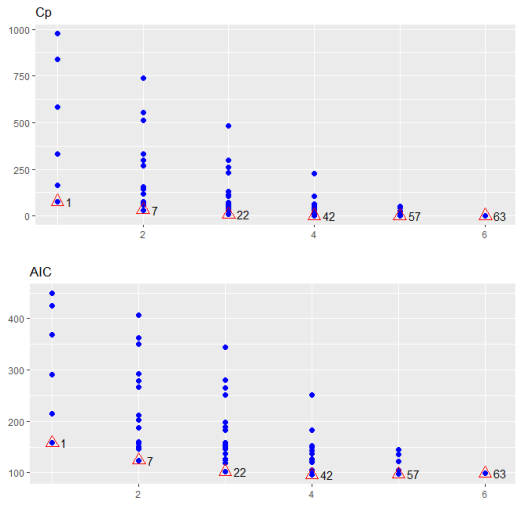
Parâmetros	Estimativa	Erro Padrão	Estatística t	P-valor
β_0	12,1322	0,0490	247,74	0,0000
β_1	0,0067	0,0008	8,08	0,0000
β_2	0,2578	0,0376	6,85	0,0000
β_3	0,1923	0,0369	5,21	0,0000
β_4	0,1341	0,0450	2,98	0,0032

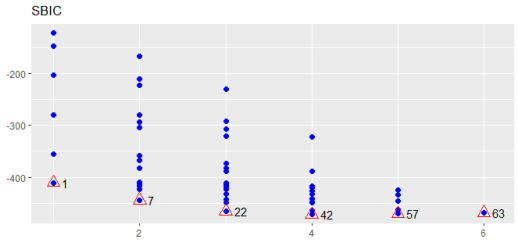
Observando a Tabela 19, nota-se que todas as variáveis são significantes, pois apresentaram um p-valor menor que o α, ou seja, foram retiradas as variáveis Banheiro e Suíte. Dessa forma, tem-se o Modelo 2 final, e a próxima etapa é verificar se os modelos selecionados pelos critérios de seleção e pelos métodos automáticos convergem para este mesmo modelo.

Modelo selecionado pelos critérios de seleção

Nesta etapa, realiza-se a mesma análise feita sobre o modelo 1 na seção 4.3.1, aonde o objetivo é verificar se o modelo selecionado pelos critérios de seleção converge para o mesmo Modelo 2 final. Dessa forma, faz-se a análise da Figura 13, que mostra qual a melhor quantidade de variáveis, selecionada por cada critério de seleção, para compor o modelo.

Figura 13: Gráficos dos critérios de seleção





Observando a Figura 13, percebe-se que as melhores quantidades para compor o modelo são 3, 4 e 5, ambas estão com valores próximos. Porém, em alguns gráficos, apesar da pouca diferença, é possível notar que o número 4 possui as menores medidas e, assim, pode-se dizer que esta é a melhor escolha.

Além disso, por meio de alguns critérios de seleção, também é possível saber qual o melhor modelo para cada quantidade de variáveis selecionada anteriormente. A partir disso, foi feito uma tabela com o *ranking* dos três melhores modelos com 3, 4 e 5 variáveis, como é mostrado na Tabela 20.

Tabela 20: *Ranking* dos três melhores modelos com 3, 4 e 5 variáveis

Ranking	Nº de variáveis	Bairro	Área	Quarto	Banheiro	Suíte	Vaga	R^2	R^2_a	C_p	BIC
1º	3	0	1	1	0	0	1	0,84	0,84	10,46	-346,46
2º	3	1	1	1	0	0	0	0,83	0,82	28,59	-329,30
3º	3	0	1	1	1	0	0	0,82	0,82	34,79	-323,76
1º	4	1	1	1	0	0	1	0,85	0,84	3,63	-350.08
2º	4	0	1	1	0	1	1	0,84	0,84	11,41	-342,20
3º	4	0	1	1	1	0	1	0,84	0,84	12,44	-341,18
1º	5	1	1	1	0	1	1	0,85	0,84	5,12	-345,31
2º	5	1	1	1	1	0	1	0,85	0,84	5,63	-344,78
3º	5	0	1	1	1	1	1	0,84	0,84	13,32	-336,99

A partir dos resultados da Tabela 20, observa-se que os melhores modelos com 3, 4 e 5 variáveis, respectivamente, são:

$$Log(Valor)=\beta_0+\beta_1 Area+\beta_2 Quarto+\beta_3 Vaga+Erro$$

$$Log(Valor)=\beta_0+\beta_1 Area+\beta_2 Quarto+\beta_3 Vaga+\beta_4 Bairro_{(asasul)}+Erro$$

$$Log(Valor)=\beta_0+\beta_1 Area+\beta_2 Quarto+\beta_3 Vaga+\beta_4 Bairro_{(asasul)}+\beta_5 Suite+Erro$$

Além disso, comparando os três modelos acima em relação aos critérios C_p e BIC, percebe-se que o modelo com 4 variáveis possui as menores medidas, ou seja, ele é melhor que os modelos com 3 e 5 variáveis, convergindo, assim, para o mesmo Modelo 2 final.

Modelo selecionado pelos métodos automáticos

Nesta etapa, realiza-se a mesma análise feita sobre o modelo 1 na seção 4.3.1, aonde o objetivo é verificar se o modelo selecionado pelos métodos automáticos converge para o mesmo Modelo 2 final, da mesma forma como o modelo de seleção de variáveis convergiu. Dessa maneira, com o uso da plataforma RStudio, foi constatado que ambos os métodos convergiram para o mesmo modelo, que foi:

$$Log(Valor)=\beta_0+\beta_1 Area+\beta_2 Quarto+\beta_3 Vaga+\beta_4 Bairro_{(asasul)}+Erro$$

Nesse sentido, percebe-se que o modelo selecionado pelos métodos automáticos convergiu para o mesmo Modelo 2 final.

Portanto, pode-se dizer que, considerando um modelo aonde a variável resposta está na escala logarítmica e com todas variáveis explicativas na sua escala original, o modelo mais adequado é o modelo composto por Área, Quarto, Vaga e Bairro. No entanto, ainda é necessário realizar a análise de diagnóstico deste modelo para examinar se este atende a todos os pressupostos.

Análise de diagnóstico do Modelo 2 final

Primeiramente, realiza-se a verificação dos três pressupostos:

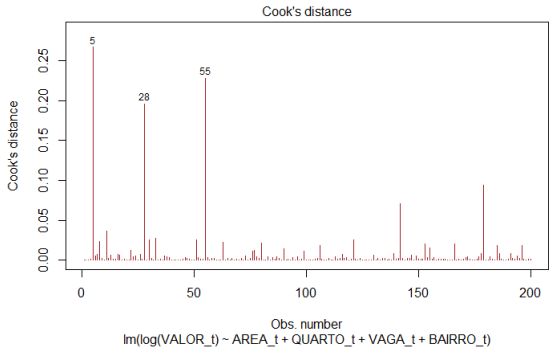
Tabela 21: Resultado dos Testes

Pressupostos	Testes	P-valor
Normalidade	Shapiro-Wilk	$< 0{,}001$
Independência dos erros	Durbin-Watson	0,8559
Homogeneidade da variância	Breusch-Pagan	0,5115

A partir dos resultados da Tabela 21, nota-se que, considerando um nível de significância de 5%, o teste de normalidade resultou em um p-valor menor que o α, ou seja, rejeitou sua hipótese nula. Assim, há evidências para se dizer que os dados não possuem distribuição normal. Tal fato pode ser justificado pela presença de *outliers* no conjunto de dados e, assim, mais a frente será construído um novo modelo sem esses valores discrepantes e será verificado novamente estes pressupostos.

Outrossim, é importante examinar a presença de observações influentes. Nesse sentido, para esse análise, utiliza-se a mesma técnica que foi aplicada sobre o modelo 1 na seção 4.3.1, chamada DFCOOKS. Dessa forma, faz-se o estudo do gráfico abaixo:

Figura 14: Gráfico para verificar observações influentes



Analisando a Figura 14, é possível notar que existem cinco observações possivelmente influentes, que são: 5, 28, 55, 142 e 179. Além disso, é importante observar que há outras observações com essa possibilidade, mas que não foram destacadas no gráfico como três das citadas anteriormente, por exemplo: 8, 11, 30, 33, 51, 63, 80, 121, 153 e 166, totalizando, assim, em 14.

Ademais, é válido analisar também se existe multicolinearidade, já que foi visto na análise bidimensional que Área e Quarto possuem alta correlação. Dessa forma, calcula-se os VIF_k's de cada variável do modelo, cujos resultados foram:

Tabela 22: Resultado dos VIF_k's

Variável	VIF_k's
Área	4,195399
Quarto	3,453733
Vaga	1,565043
Bairro(asa sul)	1,063754

A partir da análise da Tabela 22, nota-se que as variáveis Vaga e Bairro possuem VIF próximo de 1, ou seja, essas variáveis não estão correlacionadas com as demais. No entanto, ao analisar as variáveis Área e Quarto, percebe-se que elas tiveram valores de VIF consideralvemente maiores que um, indicando que elas possuem forte correlação entre si. No entando, sabendo que o valor máximo do VIF (4,20) é menor que 10, e analisando o VIF médio, cujo valor é de 2,57, ou seja, não é um valor consideravelmente maior que 1, pode-se dizer que a multicolinearidade não está influenciando as estimativas dos parâmetros.

Construção do modelo sem os valores discrepantes e influentes

Como foi visto na análise de diagnóstico, observou-se que este modelo rejeitou o pressuposto de normalidade, fato este provavelmente causado pela presença de valores

discrepantes e influentes. Assim, decidiu-se tirar todas essas observações, para verificar se, sem esses dados, o modelo atende a todos os pressupostos. Dessa forma, realizando novamente os testes, obteve-se os seguintes resultados:

Tabela 23: Resultados dos testes

Teste	P-valor
Shapiro-Wilk	0,2344
Durbin-Watson	0,7885
Breusch-Pagan	0,0623

A partir da análise da Tabela 23, percebe-se que todos os testes resultaram em um p-valor maior que o nível de significância α, cujo valor é de 5%. Assim, pode-se dizer que não há evidências suficientes para rejeitar as hipóteses nulas de cada teste, corroborando com a ideia de que esses dados estavam prejudicando o ajuste do modelo. Portanto, conclui-se que este modelo, após tirar todos os valores discrepantes e influentes, atendeu a todos os pressupostos e, assim, é correto dizer que ele é o mais adequado para essa estrutura.

4.3.3 Modelo de Regressão 3

Nessa seção, será realizada a análise do modelo completo categorizado. Este, por sua vez, será formado por: Área, Quarto, Banheiro, Suíte, Vaga e Bairro, como sendo as variáveis explicativas ; Valor como sendo a variável resposta, em sua escala original. Porém, as variáveis Quarto, Banheiro, Suíte e Vaga não serão empregadas como numéricas, tal como foi feito nos modelos 1 e 2, e sim como categóricas, ou seja, elas serão divididas em duas categorias, da maneira como foi citado na seção 4.1. Dessa forma, o modelo tem a seguinte estrutura:

$$Valor=\beta_0+\beta_1 Area+\beta_2 Quarto_{(>2)}+\beta_3 Banheiro_{(>3)}+\beta_4 Suite_{(tem)}+\beta_5 Vaga_{(tem)}+\beta_6 Bairro_{(asasul)}+Erro$$

Por conseguinte, com o objetivo de verificar quais variáveis são significativas, aplica-se o teste t-Student sobre o modelo acima. Nesse sentido, todas as variáveis que possuírem um p-valor maior que o nível de significância, cujo valor é de 5%, serão retiradas do modelo. Assim, após retirar todas as variáveis que não foram significantes, chegou-se no seguinte modelo:

$$Valor=\beta_0+\beta_1 Area+\beta_2 Banheiro_{(>3)}+\beta_3 Vaga_{(tem)}+Erro$$

Tabela 24: Análise do Modelo 3

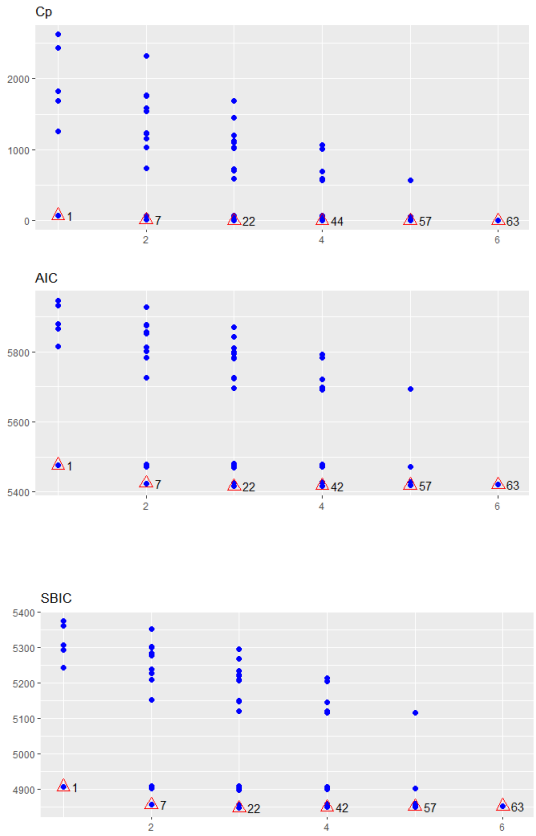
Parâmetros	Estimativa	Erro Padrão	Estatística t	P-valor
β_0	-153258,1890	25942,0519	-5,91	0,0000
β_1	10994,7932	321,4457	34,20	0,0000
β_2	210665,4855	63665,3589	3,31	0,0011
β_3	231849,2360	28701,3942	8,08	0,0000

Observando a Tabela 24, nota-se que todas as variáveis são significantes, pois apresentaram um p-valor menor que o α, ou seja, foram retiradas as variáveis Bairro, Quarto e Suíte. Dessa forma, tem-se o Modelo 3 final, e a próxima etapa é verificar se os modelos selecionados pelos critérios de seleção e pelos métodos automáticos convergem para este mesmo modelo.

Modelo selecionado pelos critérios de seleção

Nesta etapa, realiza-se a mesma análise feita sobre o modelo 1 na seção 4.3.1, aonde o objetivo é verificar se o modelo selecionado pelos critérios de seleção converge para o mesmo Modelo 3 final. Dessa forma, faz-se a análise da Figura 15, que mostra qual a melhor quantidade de variáveis, selecionada por cada critério de seleção, para compor o modelo.

Figura 15: Gráficos dos critérios de seleção



Observando a Figura 15, percebe-se que as melhores quantidades para compor o modelo são 2, 3, 4 e 5, ambas estão com valores próximos. Porém, as quantidades 3 e 4, apesar da pouca diferença, possuem as menores medidas e, entre elas, é capaz de notar que a quantidade 3 é a menor delas. Assim, pode-se dizer que esta é a melhor escolha.

Além disso, por meio de alguns critérios de seleção, também é possível saber qual o melhor modelo para cada quantidade de variáveis selecionada anteriormente. A partir disso, foi feito uma tabela com o *ranking* dos três melhores modelos com 3 e 4 variáveis, como é apresentado na Tabela 25.

Tabela 25: *Ranking* dos três melhores modelos com 3 e 4 variáveis

Ranking	Nº de variáveis	Área	Quarto(>2)	Banheiro(>3)	Suíte(tem)	Vaga(tem)	Bairro(asa sul)	R^2	R_a^2	C_p	BIC
1º	3	1	0	1	0	1	0	0	0,93	1,90	-522,48
2º	3	1	1	0	0	1	0	0	0,93	10,75	-513,55
3º	3	1	0	0	1	1	0	0	0,93	12,21	-512,11
1º	4	1	0	1	1	1	0	0	0,93	3,44	-517,65
2º	4	1	1	1	0	1	0	0	0,93	3,71	-517,37
3º	4	1	0	1	0	1	1	0	0,93	3,73	-517,35

A partir dos resultados da Tabela 25, observa-se que os melhores modelos com 3 e 4 variáveis, respectivamente, são:

$$Valor=\beta_0+\beta_1 Area+\beta_2 Banheiro_{(>3)}+\beta_3 Vaga_{(tem)}+Erro$$

$$Valor=\beta_0+\beta_1 Area+\beta_2 Banheiro_{(>3)}+\beta_3 Suite_{(tem)}+\beta_4 Vaga_{(tem)}+Erro$$

Além disso, comparando os dois modelos acima em relação aos critérios C_p e BIC, percebe-se que o modelo com 3 variáveis possui as menores medidas, ou seja, ele é melhor que o modelo com 4 variáveis, corroborando com o que foi dito anteriormente e convergindo, assim, para o mesmo Modelo 3 final.

Modelo selecionado pelos métodos automáticos

Nesta etapa, realiza-se a mesma análise feita sobre o modelo 1 na seção 4.3.1, aonde o objetivo é verificar se o modelo selecionado pelos métodos automáticos converge para o mesmo Modelo 3 final, da mesma forma como o modelo de seleção de variáveis convergiu. Dessa maneira, com o uso da plataforma RStudio, foi constatado que ambos os métodos convergiram para o mesmo modelo, que foi:

$$Valor=\beta_0+\beta_1 Area+\beta_2 Banheiro_{(>3)}+\beta_3 Vaga_{(tem)}+Erro$$

Nesse sentido, percebe-se que o modelo selecionado pelos métodos automáticos convergiu para o mesmo Modelo 3 final.

Portanto, pode-se dizer que, considerando um modelo sem transformação na variável resposta, com Quarto, Banheiro, Suíte e Vaga na escala categórica, e Área e Bairro na escala original, o modelo mais adequado é o modelo composto por Área, Banheiro e Vaga. No entanto, ainda é necessário realizar a análise de diagnóstico deste modelo para examinar se este atende a todos os pressupostos.

Análise de diagnóstico do Modelo 3 final

Primeiramente, realiza-se a verificação dos três pressupostos:

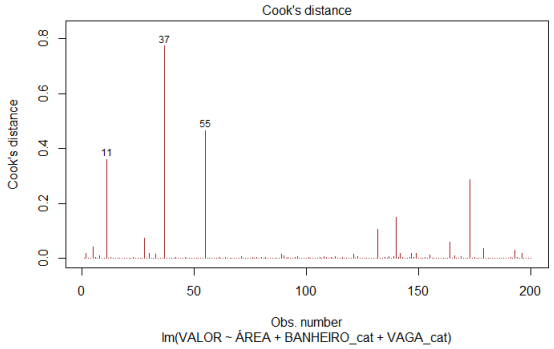
Tabela 26: Resultado dos Testes

Pressupostos	Testes	P-valor
Normalidade	Shapiro-Wilk	$< 0{,}001$
Independência dos erros	Durbin-Watson	0,7219
Homogeneidade da variância	Breusch-Pagan	$< 0{,}001$

A partir dos resultados da Tabela 26, nota-se que, considerando um nível de significância de 5%, os testes de normalidade e homogeneidade da variância resultaram em um p-valor menor que o α, ou seja, rejeitaram suas respectivas hipóteses nulas. Assim, há evidências para se dizer que os dados não possuem distribuição normal e nem variâncias iguais. Tal fato pode ser justificado pela presença de *outliers* no conjunto de dados e, assim, mais a frente será construído um novo modelo sem esses valores discrepantes e será verificado novamente estes pressupostos.

Outrossim, é importante examinar a presença de observações influentes. Nesse sentido, para esse análise, utiliza-se a mesma técnica que foi aplicada sobre o modelo 1 na seção 4.3.1, chamada DFCOOKS. Dessa forma, faz-se o estudo do gráfico abaixo:

Figura 16: Gráfico para verificar observações influentes



Analisando a Figura 16, é possível notar que existem cinco observações possivelmente influentes, que são: 11, 37, 55, 140 e 173. Além disso, é importante observar que há outras observações com essa possibilidade, mas que não foram destacadas no gráfico como as cinco citadas anteriormente, por exemplo: 5, 28, 132, 164, 179 e 193, totalizando, assim, em 11.

Ademais, é válido analisar também se existe multicolinearidade, já que foi visto na análise bidimensional que Área e Banheiro possuem alta correlação. Dessa forma, calcula-se os VIF_k's de cada variável do modelo, cujos resultados foram:

Tabela 27: Resultado dos VIF_k's

Variavel	VIF_k's
Área	1,773197
Banheiro(>3)	1,507290
Vaga(tem)	1,252054

A partir da análise da Tabela 27, nota-se que ambas as variáveis possuem VIF próximo de 1, ou seja, essas variáveis não estão correlacionadas entre si. Além disso, analisando o VIF médio, cujo valor é de 1,51, percebe-se que não é um valor consideravelmente maior que 1 e, assim, pode-se dizer que a multicolinearidade não está influenciando as estimativas dos parâmetros.

Construção do modelo sem os valores discrepantes e influentes

Como foi visto na análise de diagnóstico, observou-se que este modelo rejeitou os pressupostos de normalidade e homogeneidade da variância, fato este provavelmente causado pela presença de valores discrepantes e influentes. Assim, decidiu-se tirar todas essas observações, para verificar se, sem esses dados, o modelo atende a todos os pressupostos. Dessa forma, realizando novamente os testes, obteve-se os seguintes resultados:

Tabela 28: Resultados dos testes

Teste	P-valor
Shapiro-Wilk	0,3549
Durbin-Watson	0,8909
Breusch-Pagan	0,3947

A partir da análise da Tabela 28, percebe-se que todos os testes resultaram em um p-valor maior que o nível de significância α, cujo valor é de 5%. Assim, pode-se dizer que não há evidências suficientes para rejeitar as hipóteses nulas de cada teste, corroborando com a ideia de que esses dados estavam prejudicando o ajuste do modelo. Portanto, conclui-se que este modelo, após tirar todos os valores discrepantes e influentes, atendeu a todos os pressupostos e, assim, é correto dizer que ele é o mais adequado para

essa estrutura.

4.3.4 Modelo de Regressão 4

Nessa seção, será realizada a análise do modelo completo categorizado, porém, será empregado o logarítmo na variável resposta. Nesse sentido, este modelo será formado por: Área, Quarto, Banheiro, Suíte, Vaga e Bairro, como sendo as variáveis explicativas ; Valor como sendo a variável resposta, aplicada na escala logarítmica. Porém, as variáveis Quarto, Banheiro, Suíte e Vaga não serão empregadas como numéricas, tal como foi feito nos modelos 1 e 2, e sim como categóricas, ou seja, elas serão divididas em duas categorias, da maneira como foi citado na seção 4.1. Dessa forma, o modelo tem a seguinte estrutura:

$$Log(Valor)=\beta_0+\beta_1 Area+\beta_2 Quarto_{(>2)}+\beta_3 Banheiro_{(>3)}+\beta_4 Suite_{(tem)}+\beta_5 Vaga_{(tem)}+\beta_6 Bairro_{(asasul)}+Erro$$

Por conseguinte, com o objetivo de verificar quais variáveis são significativas, aplica-se o teste t-Student sobre o modelo acima. Nesse sentido, todas as variáveis que possuírem um p-valor maior que o nível de significância, cujo valor é de 5%, serão retiradas do modelo. Assim, após retirar todas as variáveis que não foram significantes, chegou-se no seguinte modelo:

$$Log(Valor)=\beta_0+\beta_1 Area+\beta_2 Quarto_{(>2)}+\beta_3 Banheiro_{(>3)}+\beta_4 Vaga_{(tem)}+\beta_5 Bairro_{(asasul)}+Erro$$

Tabela 29: Análise do Modelo 4

Parâmetros	Estimativa	Erro Padrão	Estatística t	P-valor
β_0	12,2727	0,0467	262,52	0,0000
β_1	0,0102	0,0008	13,35	0,0000
β_2	0,2632	0,0684	3,85	0,0002
β_3	-0,2607	0,1112	-2,34	0,0201
β_4	0,2867	0,0482	-5,95	0,0000
β_5	0,1393	0,0444	3,14	0,0020

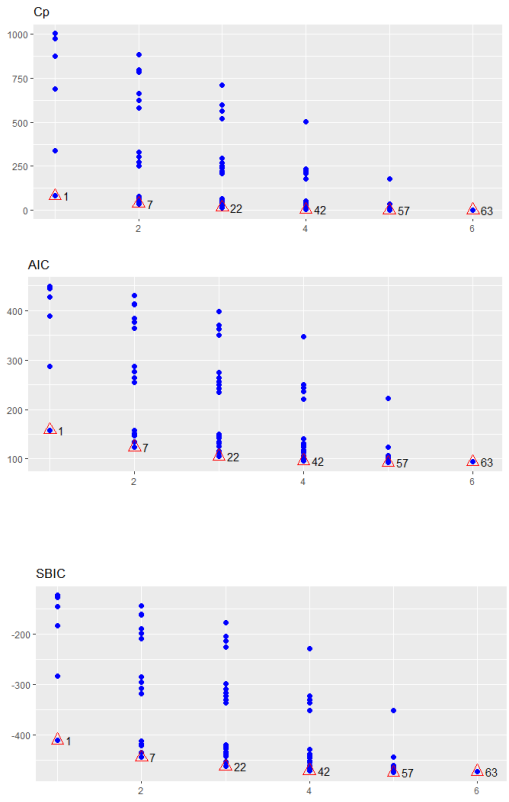
Observando a Tabela 29, nota-se que todas as variáveis são significantes, pois apresentaram um p-valor menor que o α, ou seja, foi retirada apenas a variável Suíte. Dessa forma, tem-se o Modelo 4 final, e a próxima etapa é verificar se os modelos selecionados pelos critérios de seleção e pelos métodos automáticos convergem para este mesmo

modelo.

Modelo selecionado pelos critérios de seleção

Nesta etapa, realiza-se a mesma análise feita sobre o modelo 1 na seção 4.3.1, aonde o objetivo é verificar se o modelo selecionado pelos critérios de seleção converge para o mesmo Modelo 4 final. Dessa forma, faz-se a análise da Figura 17, que mostra qual a melhor quantidade de variáveis, selecionada por cada critério de seleção, para compor o modelo.

Figura 17: Gráficos dos critérios de seleção



Observando a Figura 17, percebe-se que as melhores quantidades para compor o modelo são 4 e 5, ambas estão com valores próximos. Porém, a quantidade 5, apesar da pouca diferença, possui as menores medidas e, assim, pode-se dizer que esta é a melhor escolha.

Além disso, por meio de alguns critérios de seleção, também é possível saber qual o melhor modelo para cada quantidade de variáveis selecionada anteriormente. A partir disso, foi feito uma tabela com o *ranking* dos três melhores modelos com 4 e 5 variáveis, como é apresentado na Tabela 30.

Tabela 30: *Ranking* dos três melhores modelos com 4 e 5 variáveis

Ranking	Nº de variáveis	Área	Quarto(>2)	Banheiro(>3)	Suíte(tem)	Vaga(tem)	Bairro(asa sul)	R^2	R_a^2	C_p	BIC
1º	4	1	1	0	0	1	1	0,85	0,84	8,53	-349,73
2º	4	1	1	1	0	1	0	0,84	0,84	12,84	-345,43
3º	4	1	0	1	0	1	1	0,84	0,84	17,79	-340,61
1º	5	1	1	1	0	1	1	0,85	0,85	5,06	-350,02
2º	5	1	1	0	1	1	1	0,85	0,84	10,51	-344,45
3º	5	1	1	1	1	1	0	0,84	0,84	14,82	-340,15

A partir dos resultados da Tabela 30, observa-se que os melhores modelos com 4 e 5 variáveis, respectivamente, são:

$$Log(Valor)=\beta_0+\beta_1 Area+\beta_2 Quarto_{(>2)}+\beta_3 Vaga_{(tem)}+\beta_4 Bairro_{(asasul)}+Erro$$

$$Log(Valor)=\beta_0+\beta_1 Area+\beta_2 Quarto_{(>2)}+\beta_3 Banheiro_{(>3)}+\beta_4 Vaga_{(tem)}+\beta_5 Bairro_{(asasul)}+Erro$$

Além disso, comparando os dois modelos acima em relação aos critérios C_p e BIC, percebe-se que os valores são realmente próximos, como mostra a Figura 17, porém o modelo com 5 variáveis possui as menores medidas, ou seja, ele é melhor que o modelo com 4 variáveis, corroborando com o que foi dito anteriormente e convergindo, assim, para o mesmo Modelo 4 final.

Modelo selecionado pelos métodos automáticos

Nesta etapa, realiza-se a mesma análise feita sobre o modelo 1 na seção 4.3.1, aonde o objetivo é verificar se o modelo selecionado pelos métodos automáticos converge para o mesmo Modelo 4 final, da mesma forma como o modelo de seleção de variáveis convergiu. Dessa maneira, com o uso da plataforma RStudio, foi constatado que ambos os métodos convergiram para o mesmo modelo, que foi:

$$Log(Valor)=\beta_0+\beta_1 Area+\beta_2 Quarto_{(>2)}+\beta_3 Banheiro_{(>3)}+\beta_4 Vaga_{(tem)}+\beta_5 Bairro_{(asasul)}+Erro$$

Nesse sentido, percebe-se que o modelo selecionado pelos métodos automáticos convergiu para o mesmo Modelo 4 final.

Portanto, pode-se dizer que, considerando um modelo com transformação na variável resposta, com Quarto, Banheiro, Suíte e Vaga na escala categórica, e Área e Bairro na escala original, o modelo mais adequado é o modelo composto por Área, Quarto,

Banheiro, Vaga e Bairro. No entanto, ainda é necessário realizar a análise de diagnóstico deste modelo para examinar se este atende a todos os pressupostos.

Análise de diagnóstico do Modelo 4 final

Primeiramente, realiza-se a verificação dos três pressupostos:

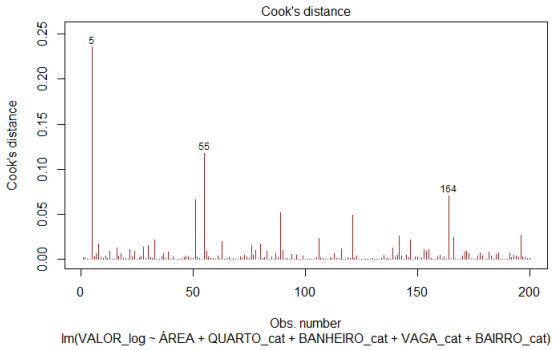
Tabela 31: Resultado dos Testes

Pressupostos	Testes	P-valor
Normalidade	Shapiro-Wilk	$< 0{,}001$
Independência dos erros	Durbin-Watson	0,636
Homogeneidade da variância	Breusch-Pagan	0,4053

A partir dos resultados da Tabela 31, nota-se que, considerando um nível de significância de 5%, o teste de normalidade resultou em um p-valor menor que o α, ou seja, rejeitou sua hipótese nula. Assim, há evidências para se dizer que os dados não possuem distribuição normal. Tal fato pode ser justificado pela presença de *outliers* no conjunto de dados e, assim, mais a frente será construído um novo modelo sem esses valores discrepantes e será verificado novamente estes pressupostos.

Outrossim, é importante examinar a presença de observações influentes. Nesse sentido, para esse análise, utiliza-se a mesma técnica que foi aplicada sobre o modelo 1 na seção 4.3.1, chamada DFCOOKS. Dessa forma, faz-se o estudo do gráfico abaixo:

Figura 18: Gráfico para verificar observações influentes



Analisando a Figura 18, é possível notar que existem três observações possivelmente influentes, que são: 5, 55 e 164. Além disso, é importante observar que há outras observações com essa possibilidade, mas que não foram destacadas no gráfico como as citadas anteriormente, por exemplo: 33, 51, 89, 106, 121, 142, 147, 166 e 196, totalizando, assim, em 12.

Ademais, é válido analisar também se existe multicolinearidade, já que foi visto na análise bidimensional que Área, Quarto e Banheiro possuem alta correlação. Dessa forma, calcula-se os VIF_k's de cada variável do modelo, cujos resultados foram:

Tabela 32: Resultado dos VIF_k's

Variavel	VIF_k's
Área	3,657801
Quarto(>2)	2,552197
Banheiro(>3)	1,681834
Vaga(tem)	1,289594
Bairro(asa sul)	1,060695

A partir da análise da Tabela 32, nota-se que as variáveis Banheiro, Vaga e Bairro possuem VIF próximo de 1, ou seja, essas variáveis não estão correlacionadas com as demais. No entanto, ao analisar as variáveis Área e Quarto, percebe-se que elas tiveram valores de VIF consideralvemente maiores que um, indicando que elas possuem forte correlação entre si. No entando, sabendo que o valor máximo do VIF (3,66) é menor que 10, e analisando o VIF médio, cujo valor é de 2,05, ou seja, não é um valor consideravelmente maior que 1, pode-se dizer que a multicolinearidade não está influenciando as estimativas dos parâmetros.

Construção do modelo sem os valores discrepantes e influentes

Como foi visto na análise de diagnóstico, observou-se que este modelo rejeitou o pressuposto de normalidade, fato este provavelmente causado pela presença de valores discrepantes e influentes. Assim, decidiu-se tirar todas essas observações, para verificar se, sem esses dados, o modelo atende a todos os pressupostos. Dessa forma, realizando novamente os testes, obteve-se os seguintes resultados:

Tabela 33: Resultados dos testes

Teste	P-valor
Shapiro-Wilk	0,1496
Durbin-Watson	0,7259
Breusch-Pagan	0,01168

A partir da análise da Tabela 33, percebe-se que, com a saída dos valores discrepantes e influentes, o problema da normalidade foi corrigido, pois resultou em um p-valor maior que o nível de significância α, cujo valor é de 5%. Porém, nota-se que o pressuposto de homogeneidade da variância passou a ser rejeitado, dado que obteve um p-valor menor que o nível de significância, rejeitando, assim, a hipótese nula. Portanto, conclui-se que este modelo, mesmo após tirar todos os valores discrepantes e influentes, não atendeu a todos os pressupostos e, assim, é correto dizer que ele é não é o mais adequado para essa

estrutura, ainda que os critérios de seleção e os métodos automáticos tenham convergido para este modelo.

No entanto, anteriormente, quando foi analisado qual o melhor modelo com base nos critérios de seleção, percebeu-se que o modelo com 4 variáveis apresentou valores bastante próximos do modelo com 5 variáveis. Dessa forma, realiza-se, agora, a mesma análise feita acima sobre este modelo, com intuito de verificar se ele atende a todos os pressupostos. Nesse sentido, observa-se abaixo a estrutura desse modelo e os resultados dos testes respectivamente:

$$Log(Valor)=\beta_0+\beta_1 Area+\beta_2 Quarto_{(>2)}+\beta_3 Vaga_{(tem)}+\beta_4 Bairro_{(asasul)}+Erro$$

Tabela 34: Resultados dos testes

Teste	P-valor
Shapiro-Wilk	0,9615
Durbin-Watson	0,4728
Breusch-Pagan	0,1548

A partir da análise da Tabela 34, percebe-se que todos os testes resultaram em um p-valor maior que o nível de significância α, cujo valor é de 5%. Assim, pode-se dizer que não há evidências suficientes para rejeitar as hipóteses nulas de cada teste. Portanto, conclui-se que este modelo, após tirar todos os valores discrepantes e influentes, atendeu a todos os pressupostos e, assim, é correto dizer que ele é o mais adequado para essa estrutura, mesmo que os critérios de seleção e os métodos automáticos não tenham convergido para este modelo, sendo, dessa forma, definido como o Modelo 4 final.

4.3.5 Análise de desempenho dos modelos de regressão

Nessa seção, os quatro modelos de regressão finais selecionados anteriormente serão comparados, com o intuito de descobrir qual deles possui a melhor capacidade preditiva. Nesse sentido, para realizar essa análise faz-se o uso de algumas medidas, como o PRESS (Soma de Quadrados do Erro Residual Previsto), o MSPR (Erro Quadrático Médio de Previsão), o R^2 e o R_a^2. Ainda em relação à base de treino, compara-se esses modelos em relação ao PRESS, como mostra a tabela abaixo:

Tabela 35: Comparação dos Modelos Finais - Base de treinamento

Medida	Modelo 1 Final	Modelo 2 Final	Modelo 3 Final	Modelo 4 Final
PRESS	220,77	209,98	217,41	207,36

Observando a Tabela 35, nota-se que o modelo 4 final apresentou o menor valor de PRESS, indicando que este modelo, em relação a essa medida de desempenho, possui a melhor capacidade preditiva, seguida do modelo 2 final, o qual possui um valor de PRESS bastante próximo.

Além disso, agora em relação à base de validação, compara-se os modelos em relação ao MSPR, ao R^2 e ao R^2_a, como mostra a tabela a seguir:

Tabela 36: Comparação dos Modelos Finais - Base de validação

Medidas	Modelo 1 Final	Modelo 2 Final	Modelo 3 Final	Modelo 4 Final
MSPR	0,231	0,227	0,277	0,239
R^2	0,780	0,788	0,722	0,779
R^2_a	0,776	0,782	0,717	0,773

A partir dos resultados da Tabela 36, nota-se que ambos os modelos, em todas as medidas, possuem valores bastante próximos. Porém, é possível concluir que:

- O modelo 2 final possui o menor valor de MSPR e, nesse caso, quanto menor o valor, melhor;
- O modelo 2 final possui o maior valor de R^2 e, nesse caso, quanto maior o valor, melhor;
- O modelo 2 final possui o maior valor de R^2_a e, nesse caso, quanto maior o valor, melhor.

Logo, pode-se dizer que o modelo 2 final é o que possui a melhor capacidade preditiva, pois, além de apresentar o segundo menor valor de PRESS, em relação as demais medidas ele obteve os melhores resultados, ou seja, ele foi o que melhor se adequou/ajustou ao conjunto de dados. Assim, define-se este modelo como o modelo de regressão final.

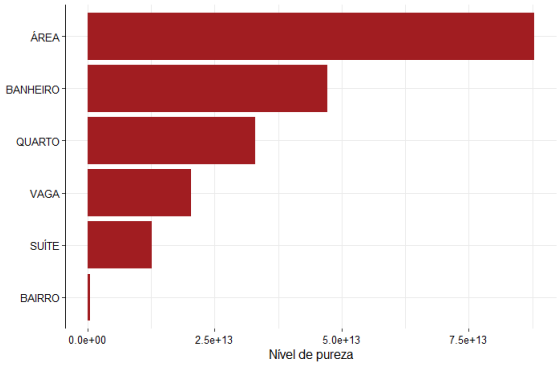
4.4 Modelos Não Paramétricos

Nessa seção, realiza-se a construção dos modelos não paramétricos, os quais serão obtidos a partir de determinados conhecimentos de aprendizagem de máquina, como: árvores de regressão, florestas aleatórias e redes neurais.

4.4.1 Modelo - Árvores de regressão

Nessa etapa, será construído o modelo de previsão a partir da metodologia de árvores de regressão. Primeiramente, deve-se determinar as divisões da árvore, e isto, por sua vez, se faz utilizando o critério de pureza, o qual consiste no erro quadrático médio (MSE). Dessa forma, o objetivo é escolher, por meio desse critério, a divisão que for mais homogênea, ou seja, mais pura para entrar na partição. Assim, calculando-se o nível de pureza de cada variável, tem-se o seguinte resultado:

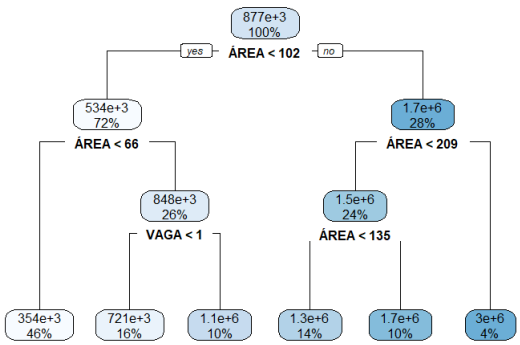
Figura 19: Gráfico do nível de pureza de cada variável



Analisando a Figura 19, percebe-se que a variável com o maior nível de pureza é a Área, portanto deve estar no topo da árvore, na raíz e, provalmente, nas demais divisões.

Além disso, outro passo importante é podar a árvore, ou seja, retirar cada nó/ramo da árvore, um por vez, e analisar o que acontece com o erro estimado no conjunto de validação. Dessa forma, aplicando esse processo na árvore, chega-se no seguinte resultado:

Figura 20: Árvore podada



Observando a Figura 20, percebe-se que:

- Como era esperado, a variável Área está no topo da árvore, pois possui o maior nível de pureza, e aparece em outras divisões;
- Apesar da variável Vaga possuir um dos menores níveis de pureza, esta aparece em uma das divisões;
- Não houve diferença na árvore após ela ser podada, evidenciando que antes a árvore não apresentava *overfitting* (variância alta).

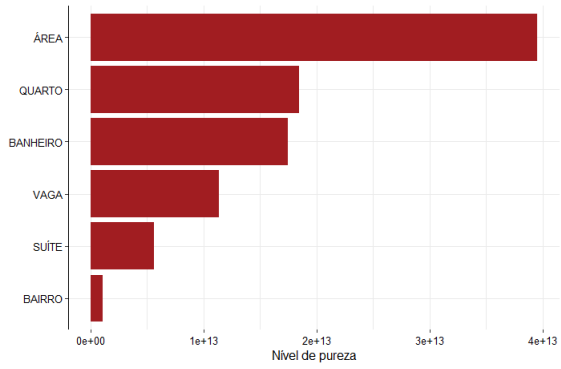
Acerca do segundo tópico, é válido ressaltar que a própria medida de pureza apresenta problemas. Se existir duas variáveis que tem uma correlação muito alta e que ao mesmo tempo são importantes, a importância dessas variáveis diminuem pela metade basicamente. Dessa forma, é fundamental que, ao aplicar e analisar essa medida, se tenha conhecimento acerca desse problema. Tal fato pode explicar o motivo pelo qual a variável Vaga aparece em uma das divisões, e Quarto e Banheiro não, pois estas são muito correlacionadas com a variável Área.

4.4.2 Modelo - Florestas Aleatórias

Nessa etapa, será construído o modelo de previsão a partir da metodologia de florestas aleatórias, que combina centenas de árvores de regressão, obtidas a partir de amostras bootstrap do banco de dados, para chegar a uma melhor previsão.

Nesse sentido, assim como é feito nas árvores de regressão, calcula-se também o nível de pureza de cada variável, com o objetivo de definir qual delas entrará nas divisões de cada árvore construída. Dessa forma, obteve-se os seguintes resultados:

Figura 21: Gráfico do nível de pureza de cada variável



Analisando a Figura **??**, percebe-se que a variável com o maior nível de pureza é a Área, da mesma forma como foi visto na Figura 19. Portanto, esta variável deve estar

no topo das árvores de regressão construídas, ou seja, na raíz. Além disso, é importante notar que, comparando com a Figura 19, observa-se que a variável Quarto passou a ter mais importância que Banheiro.

Ademais, após construir esse modelo na plataforma RStudio, constata-se as seguintes características:

Tabela 37: Características da Floresta Aleatória

Característica	Valor
Número de árvores	500
Tamanho das amostras	200
Número de variáveis independentes	6
Número máximo do nós	5
Tamanho de cada nó	2
R^2	0,925

4.4.3 Modelo - Redes Neurais

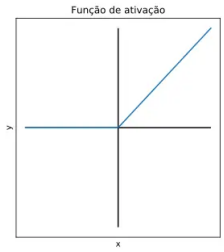
Nessa etapa, será construído o modelo de previsão a partir da metodologia de redes neurais. Dessa forma, construiu-se uma rede com as seguintes características:

- Uma camada de entrada composta por 6 neurônios, os quais correspondem às 6 variáveis explicativas do conjunto de dados: Área, Quarto, Banheiro, Suíte, Vaga e Bairro;
- Duas camadas ocultas: a primeira composta por 4 neurônios e a segunda por 3 neurônios;
- Uma camada de saída composta por um neurônio, que corresponde à variável Valor.

Além disso, é válido observar que, da camada de entrada para a primeira camada oculta, da primeira camada oculta para segunda, e da segunda camada oculta para a camada de saída foi aplicado um *dropout* de 20%, que representa uma técnica cujo objetivo é evitar ou reduzir o *overfitting*. Dessa forma, esta estrutura, no final, gerou um total de 95 parâmetros: 48 na primeira camada, 28 na segunda e 15 na terceira e 4 na última.

Ademais, outro passo importante é definir as funções de ativação que serão utilizadas na passagem de uma camada pra outra. Nesse caso, definiu-se que, da camada de entrada para a primeira camada oculta, e de uma camada oculta pra outra, seria aplicada a função de ativação ReLU (unidade linear retificada), a qual possui as seguintes características:

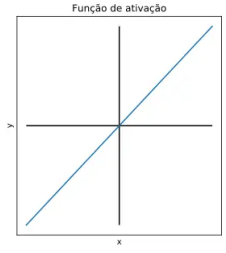
Figura 22: Gráfico da função de ativação ReLU



- Equação: A(x) = max(0,x), ou seja, ela retorna 0 para todos os valores negativos, e o próprio valor para valores positivos;
- Intervalo de valores: -[0, inf);
- Motivos da escolha: ela é não linear, ou seja, pode-se facilmente retropropagar os erros e ter várias camadas de neurônios sendo ativadas por ela ; ela consiste em uma função computacionalmente leve, pois envolve operações matemáticas mais simples ; como seu resultado é zero para valores negativos, ela tende a "apagar" alguns neurônios, isto é, apenas alguns neurônios são ativados, tornando a rede esparsa e, assim, transformando-a em uma função eficiente e fácil de calcular, o que aumenta a velocidade do treinamento; é a função de ativação mais amplamente usada, principalmente implementada em camadas ocultas de rede neural.

Porém, a função ReLU não costuma ser utilizada na camada de saída. Nesse sentido, da segunda camada oculta para a camada de saída aplicou-se a função de ativação linear, cujas características são:

Figura 23: Gráfico da função de ativação Linear



- Equação: y = ax;
- Intervalo de valores: -inf a +inf;
- Motivos da escolha: essa função pode ser utilizada em problemas de regressão, já que produz resultados em todo o domínio dos números reais, e ela é usada em apenas um lugar, na camada de saída.

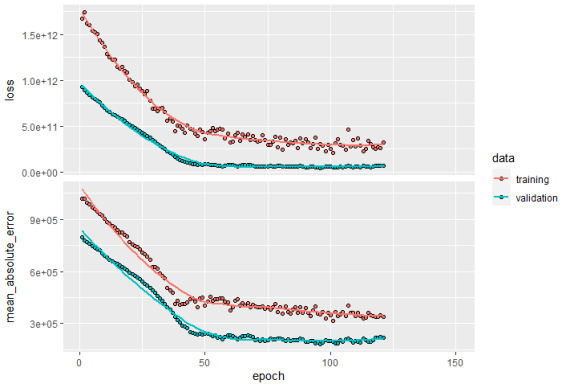
Após definir as funções de ativação, a próxima etapa consiste em estimar os betas, etapa esta que foi realizada por meio do Gradiente Descendente Estocástico. Esta técnica, por sua vez, faz algum tipo de regularização na função a qual está sendo estimada, assim, quando a rede neural é treinada por ela, o problema de *overfitting*, que acontece com o gradiente descendente normal, é muito menor, melhorando o desempenho, em termos de poder preditivo, da função estimada.

Além disso, é válido citar outros procedimentos usados, como:

- O *earling-stop*, isto é, um método para parar mais cedo o aprendizado, então as iterações não são aplicadas até minimizar a função de fato. Tal processo é utilizado, pois vai chegar uma hora em que os dados serão interpolados perfeitamente, isto é, o erro vai ser igual a zero, situação esta que aparenta ser boa, já que a minimização foi alcançada. No entanto, sabe-se que isso no geral não é bom, pois muitas vezes um estimador que interpola os dados, consequentemente, vai ter erro preditivo muito alto;
- O MSE (Erro Quadrático Médio) e erro absoluto como medidas para avaliar o desempenho dessa rede.

Agora, considerando um *epochs* de 150 e um *batch size* de 200, efetua-se o treinamento desse rede, cujo resultado foi:

Figura 24: Histórico da rede neural



Analisando a Figura 24, nota-se a presença de dois gráficos, aonde o primeiro representa o histórico da rede neural em relação ao MSE, e o segundo em relação ao erro absoluto. Além disso, cada gráfico possui duas curvas, uma delas corresponde ao conjunto de treino (80% das observações), e a outra ao conjunto de validação (20% das observações). Nesse sentido, percebe-se que, ambas as curvas, como já era esperado, se

aproximam de zero a medida que o número de treinamento/iterações aumenta. Além disso, é válido notar que, devido à aplicação do *earling-stop*, foram realizados apenas 121 *epochs* dos 150.

4.5 Análise de desempenho de cada metodologia

Nessa seção, o objetivo é comparar os modelos construídos em cada metodologia em relação ao risco preditivo (MSPR). Dessa forma, analisa-se a seguinte tabela:

Tabela 38: Capacidade preditiva de cada Metodolgia

Metodologia	Risco Preditivo (MSPR)
Modelo de Regressão Final	0,227
Modelo - Árvore de regressão	0,327
Modelo - Floresta Aleatória	0,215
Modelo - Rede Neural	0,182

Analisando a Tabela 38, percebe-se que o modelo construído por Rede Neural obteve o menor risco preditivo, com um valor de 0,182, seguido do modelo elaborado por *Random Forest* e, depois, por regressão linear, cujos valores foram, respectivamente, 0,215 e 0,227. Dessa maneira, o modelo construído por árvores de regressão foi o que apresentou o pior risco, cujo valor foi de 0,327, pois as árvores de regressão são bastante simples para apresentarem um bom poder preditivo.

Assim, é possível concluir que a metodologia a qual obeteve a melhor capacidade de predição, ou seja, que adquiriu o melhor ajuste/acurácia, foi a Rede Neural.

Nesse sentido, se sua intenção é unicamente obter um modelo com a melhor capacidade preditiva, aconselha-se a usar a metodologia de Redes Neurais. Porém, no ramo imobiliário, na maioria das vezes, as pessoas que vão vender ou comprar o imóvel desejam obter outras informações, além do valor correto do imóvel, como ver a influência de cada variável do modelo, ou ajustar o preço de acordo com alguma mudança no imóvel em questão, seja na área, na localização, no número de quartos, entre outras características.

Assim, quando se leva em consideração esses outros pontos além da predição, não é aconselhável usar os métodos de *Machine Learning*, pois este, por sua vez, são muito abstratos, são como um caixa-preta, não se sabe o que está acontecendo dentro daquele algoritmo que está sendo calculado, não se tem muito controle da situação. Dessa forma, pode-se dizer que eles não são explicáveis, e por isso é difícil entender como ou por que eles chegaram a uma determinada decisão, ou seja, significa que o modelo construído simplesmente precisa ser confiável como está e os resultados aceitos como estão.

Portanto, apesar do modelo de regressão ter apresentado uma capacidade preditiva pior que os modelos do *Random Forest* e da Rede Neural, seu desempenho não foi muito abaixo, evidenciando que esse modelo, dependendo da situação e das informações que se desejam, leva vantagens e é mais relevante que os demais métodos. Nesse sentido, pode-se dizer que o modelo de regressão nos traz determinados benefícios os quais os modelos de *Machine Learning* não possuem ou não conseguem obter com facilidade por exigir mais elaboração computacionalmente, como:

- Ver a influência direta de cada variável explicativa em relação à variável resposta, ou seja, se aumentar o tamanho da área, ou diminuir o número de banheiros por exemplo, verificar o que isso provocaria no valor final;
- Fazer uma estimativa intervalar sem muita dificuldade;
- Ter uma capacidade linear direta, o que também é uma vantagem dos modelos de regressão.

5 Considerações Finais

Diante dos resultados gerados durante o desenvolvimento do relatório, conclui-se que, comparando os modelos construídos em cada metodologia aplicada (Regressão Linear, Árvores de Regressão, Floresta Aleatória e Redes Neurais) apenas em relação ao seu risco preditivo (MSPR), as metodologias de *Machine Learning*, com exceção das árvores de regressão, obtiveram modelos com capacidades preditivas melhores que os modelos de regressão. Porém, apesar da pouca diferença, o modelo construído por Rede Neural adquiriu um erro preditivo menor que o construído por *Random Forest*, sendo, assim, a metodologia com a melhor capacidade de predição.

Dessa maneira, dado que o intuito final deste trabalho é propor uma solução para corrigir as falhas que ainda existem no ramo imobiliário em relação à precificação dos imóveis (avaliações acima ou abaixo do mercado), ou seja, encontrar a metodologia estatística mais adequada e precisa para se prever o valor de um imóvel, pode-se dizer que o objetivo foi alcançado.

Portanto, visando obter uma melhoria sobre o mercado imobiliário, através de avaliações corretas, o que fará com que o mercado possua a dinâmica e a liquidez esperada, recomenda-se utilizar a metodologia de Redes Neurais. Porém, é importante frisar que no ramo imobiliário pode haver outros interesses além de obter uma boa predição e de ter velocidade no processamento. Nesse casos, percebeu-se que as técnicas de regressão são mais vantajosas, mais relevantes e, além disso, apresentam mais facilidade, tanto no quesito programacional quanto no de interpretabilidade, que as demais metodologias.

Referências

AMARY, F. *A importância e a responsabilidade de uma avaliação imobiliária*. 2014. Acesso em 26 de fevereiro de 2022. Disponível em: <https://www2.jornalcruzeiro.com.br/materia/533688/a-importancia-e-a-responsabilidade-de-uma-avaliacao-imobiliaria>.

BANIB. *Por que a avaliação de imóveis é importante?* 2019. Acesso em 26 de fevereiro de 2022. Disponível em: <https://blog.banib.com/avaliacao-de-imoveis/>.

BUSSAB WILTON DE O. MORETTIN, P. A. *Estatística Básica*. [S.l.]: Saraiva, 2018.

Colaboradores da Wikipédia. *Método dos mínimos quadrados*. 2022. Último acesso em 26 de fevereiro de 2022. Disponível em: <https://pt.wikipedia.org/wiki/M\%C3\%A9todo_dos_m\%C3\%ADnimos_quadrados>.

COSSI, M. L. M. L. F. R. C. A. M. *Avaliação do modelo de regressão linear múltipla e redes neurais artificiais na previsão do ganho de massa em animais*. Tese (Doutorado) — Universidade Estadual Paulista, Faculdade de Engenharia de Ilha Solteira, Departamento Matemática, 2017.

EXAME. *O futuro do mercado imobiliário em um mundo pós-pandemia*. 2021. Acesso em 26 de fevereiro de 2022. Disponível em: <https://exame.com/colunistas/genoma-imobiliario/o-futuro-do-mercado-imobiliario-em-um-mundo-pos-pandemia/>.

IZBICKI, R.; SANTOS, T. M. dos. *Aprendizado de máquina: uma abordagem estatística*. [S.l.: s.n.], 2020. ISBN 978-65-00-02410-4.

MINITAB. *Basta! Lidando com a multicolinearidade na análise de regressão*. 2019. Último acesso em 26 de fevereiro de 2022. Disponível em: <https://blog.minitab.com/pt/basta-lidando-com-a-multicolinearidade-na-analise-de-regressao>.

MQL5. *FLORESTA DE DECISÃO ALEATÓRIA NA APRENDIZAGEM POR REFORÇO: a descrição abstrata do algoritmo floresta aleatória*. 2018. Acesso em: 23 set. 2022. Disponível em: <https://www.mql5.com/pt/articles/3856>.

PEREIRA J. C.; GARSON, S. A. E. G. Construção de um modelo para o preço de venda de casas residenciais na cidade de sorocaba-sp. *GEPROS*, 2012.

PROTEL. *Entenda como funciona e qual o objetivo da avaliação de imóveis*. 2018. Acesso em 26 de fevereiro de 2022. Disponível em: <https://www.protel.com.br/protel_wp/matriadoboletim/entenda-como-funciona-e-qual-o-objetivo-da-avaliacao-de-imoveis/>.

ROQUE, R. D. C. *Estudo sobre a empregabilidade da previsão do índice BOVESPA usando Redes Neurais Artificais*. Tese (Doutorado) — Universidade Federal do Rio de Janeiro, 2009.

SAKURAI, R. *Decision Tree: Aprendendo a classificar flores do tipo Iris*. 2018. Acesso em: 23 set. 2022. Disponível em: <https://www.sakurai.dev.br/classificacao-iris/>.

SILVA, T. M. D. C. *Um estudo comparativo entre algorítmos de aprendizagem de máquina supervisionados para predição de solução de reclamações no PROCON*. Tese (Doutorado) — Centro Universitário Christus Sistemas de informação, 2021.

SOEDIL. *Vale a pena investir em imóveis em 2021?* 2021. Acesso em 26 de fevereiro de 2022. Disponível em: <https://www.soedil.com.br/blog/vale-a-pena-investir-em-imoveis-em-2021#:~:text=Com\%20certeza\%20vale\%20a\%20pena,deste\%20\%C3\%A9\%20uma\%20boa\%20escolha.>

TRYBE. *Regressão Linear Simples: O que é e como fazer?* 2022. Acesso em 26 de fevereiro de 2022. Disponível em: <https://blog.betrybe.com/regressao-linear-simples/>.

WIDESYS. *O futuro do mercado imobiliário pós-pandemia*. 2021. Acesso em 26 de fevereiro de 2022. Disponível em: <https://widesys.com.br/0-futuro-mercado-imobiliario-pos-pandemia/>.

Wikipedia contributors. *Soma dos Quadrados do Erro Residual Previsto (PRESS)*. 2022. Último acesso em 26 de fevereiro de 2022. Disponível em: <https://en.wikipedia.org/wiki/PRESS_statistic#:~:text=In\%20statistics\%2C\%20the\%20predicted\%20residual,used\%20to\%20estimate\%20the\%20model.>