Custos das habitações da Califórnia

Letícia Garcez Corrêa da Costa Bianca Ribeiro Lima Marques
Nicolas Gabriel da Costa Simões

Universidade Federal da Paraíba - UFPB - 2020

1 Introdução

Uma das principais aplicações de Aprendizagem de Máquina é a predição. Essa tarefa tem como objetivo antecipar o resultado que um grupo terá baseado num grupo de dados que foram utilizados como treinamento.

Essa predições podem ser consideradas de dois tipos, a classificação e a regressão. Classificação têm como interesse buscar uma linha reta que consiga separar as amostras em dois grupos da melhor forma. Em outtras palavras, esse método buscará produzir um classificador que seja capaz de categorizar rótulos que sejam desconhecidos. Já a Regressão, é uma forma de prever valores futuros com base nos dados já existentes.

Este relatório tem como objetivo colocar em prática alguns dos assuntos de Aprendizagem de Máquina, especificamente a regressão linear simples, múltipla e a validação cruzada. Sabendo disso, a fim de cumprir com o objetivo, utilizou-se a base de dados referente aos custos de habitação da Califórnia, obtida por meio da plataforma Kaggle, que disponibilizou o conjunto de dados usado no segundo capítulo do livro "Aprendizado de máquina prático com Scikit-Learn e TensorFlow"cujo autor é Aurélien Géron.

Com isso, os dados contêm informações do censo de 1990 da Califórnia possuindo 20.640 instâncias e 10 atributos. Cada instância está relacionado a casas próximas agrupadas de um distrito da Califórnia e possui atributos referentes a longitude, latitude, número médio de habitações, total de cômodos e quartos, população, número de famílias, valor médio da renda populacional, valor médio das casas e proximidade do oceano.

2 Análise Exploratória dos Dados

Analisando de forma mais específica cada atributo da base de dados, observa-se que com exceção da variável referente a proximidade do oceano, todas elas são numéricas. Por meio da tabela 1 é possível verificar que, em média, as casas possuem aproximadamente 29 anos, localizadas a -119,6 de longitude e 35,63 de latitude, possuindo 2.636 cômodos, incluindo cerca de 538 quartos e custando 206.856 dólares. Cada grupo de casas é formado, em média, por 1425 pessoas, sendo cerca de 500 famílias com renda média de 3,87 dezenas de milhares dólares.

Tabela 1: Descrição dos atributos quantitativos

Variáveis	Média ± desvio padrão	Mediana [intervalo interquartílico]
Longitude	-119,6 ± 2,00	-118,50 [(-121,80) - (-118,00)]
Latitude	35,63 ± 2,14	34,26 [33,93 - 37,71]
Idade média das casas	28,64 ± 12,59	29,00 [18,00 - 37,00]
Total de cômodos	2.636 ± 2.181,62	2127 [1.448 - 3.148]
Total de quartos	537,90 ± 419,27	438,0 [297,00 - 643,20]
População	1425 ± 1.132,46	1.166 [787 - 1.725]
Famílias	499,5 ± 382,33	409,0 [280,00 - 605,00]
Renda média	3,87 ± 1,90	3,53 [2,56 - 4,74]
Custo médio	206.856 ± 115.395,60	179.700 [119.600 - 264.725]

Com relação à proximidade do oceano, a unidade de medida trata-se da localização da casa w.r.t oceano/mar. Dessa forma, verfica-se por meio da tabela 2, que a maior parte das casas estão presentes a <1H do oceano representando 44,26%. As demais se encontram distribuídas no interior, próximo a baía ou ao mar e uma minoria em ilhas, em que possui apenas cinco grupos de casas.

Tabela 2: Descrição do atributo da proximidade do oceano

Classes	n	%
<1H OCEAN	9.136	44,26
INLAND	6.551	31,74
ISLAND	5	0,02
NEAR BAY	2.290	11,09
NEAR OCEAN	2.658	12,88

Por meio dos histogramas apresentados nas figuras 1, 2 e 3 é possível ver com maior clareza o comportamento dessas variáveis e pode-se dizer que a que mais se assemelha a uma distribuição normal, trata-se da idade média das casas e renda média. Além disso, é evidente o quanto as escalas diferem bastante umas das outras.

Outro fato muito importante é que as médias não permitem uma análise clara na variabilidade. Para isso, os histogramas somado aos desvios padrões apresentados na tabela 1, mostram o quanto as médias estão sendo influenciadas pelos valores extremos.

Tendo em vista que este relatório usará a regressão para predizer valores, sabe-se o quão é importante analisar a correlação entre as variáveis. Dessa forma, ao correlacionar o custo médio das casas com os demais atributos, constatou-se que a renda média da população é a que está mais relacionada com uma correlação igual a 0,69.

Figura 1: Longitude, latitude, idade média das casas e total de cômodos

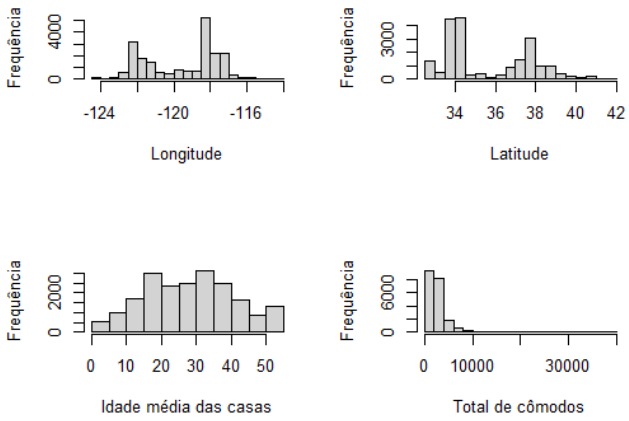


Figura 2: Total de quartos, população, famílias e renda média

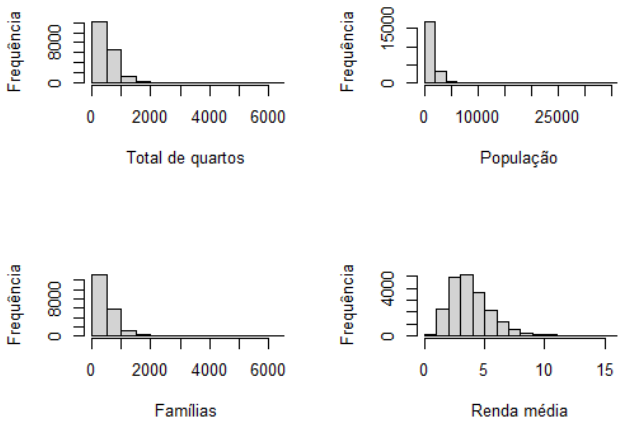
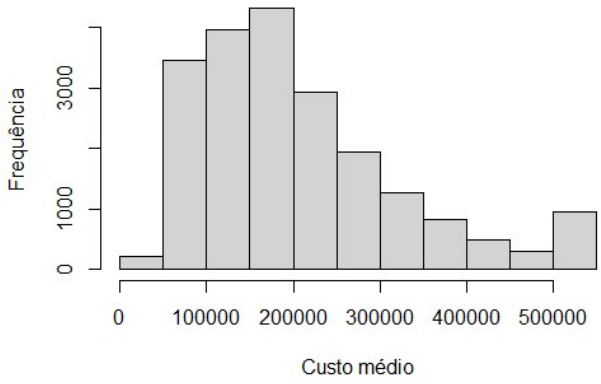


Figura 3: Custo médio das casas



3 Metodologia

Sabe-se que a base de dados precisa de um tratamento inicial antes de começar qualquer trabalho, devido a necessidade de avaliar a presença de alguma inconstância e aliar de acordo com o objetivo. O pré-processamento dos dados iniciou-se com a avaliação de 207 informações faltantes no atributo da quantidade de quartos. A fim de solucionar esse problema foram adicionadas em todas elas, a média das demais instâncias.

Por fim, assim como foi visto anteriormente, todos os atributos apresentam escalas e distribuições diferentes entre si. Logo, foi realizada a normalização de todos os atributos para a implementação do modelo.

Uma das técnicas estatísticas mais usadas para a solução de problemas reais é a regressão. Com ela é possível estudar o comportamento que um atributo tem em relação a outros que são independentes e determinam sua variabilidade. Dessa forma, o relacionamento entre essas variáveis serão descritas por uma equação matemática. Observe a figura 4 abaixo. Ela é formada pelo beta, que trata-se do coeficiente linear (beta 0) e angular (os demais), pelo x, que são as variáveis independentes que explicam y e por fim, o erro.

Vale salientar que, como será realizado um modelo de regressão linear, foi desconsiderado o atributo que não iria contribuir para a construção desse modelo. Portanto, tratando-se de uma variável categórica, as instâncias referentes à proximidade do oceano foram retiradas dos modelos.

Figura 4: Modelo de regressão linear

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

Como dito anteriormente, nesse trabalho foi feito dois tipos de regressão. Com o objetivo de estimar a variável resposta em função da variável explicativa, foi utilizado o modelo de regressão linear simples (MRLS). Esse modelo usa uma única variável explicativa, sendo apenas x1 da fórmula mostrada acima. Essa variável foi selecionada de conforme o maior valor da correlação calculado na análise descritiva dos atributos.

Portanto, o MRLS é iniciado com o gráfico de dispersão, representação gráfica da associação entre os dados podendo também analisar o tipo de correlação entre os atributos. Outras formas de verificar a correlação é fazendo o teste de correlação (Test for Association/Correlation Between Paired Samples) para confirmação do gráfico. Logo após o ajuste do modelo, é feito os testes T e F (teste de adequação global) como também a análise de variância (ANOVA), que verifica a adequação ou a qualidade do ajuste do modelo de regressão, rejeitando ou não a hipótese nula por meio do p-valor (valor significante = p-valor $<$ nível de 5% significância adotado).

Somado a isso, utilizou-se também o modelo de regressão múltipla (MRLM). Este modelo permite trabalhar com mais de uma variável explicativa (beta de 0 a p, como mostrado na figura 4) e por isso, foram englobados todos os atributos numéricos da base de dados. Esse modelo também permite a realização dos testes feitos na regressão simples. No entanto, foi acrescentado no relatório a análise residual que verifica os pres-

supostos da normalidade, homocedasticidade, linearidade e autocorrelação mais a análise de diagnósticos que indica quais os outliers que podem ser retirados do modelo (pontos aberrantes que não são de alavanca nem de influência).

Para finalizar, com o intuito de avaliar os modelos, foi realizado a validação cruzada da forma mais simples possível. A validação cruzada trata-se de uma das melhores técnicas para analisar a eficiência do modelo principalmente na predição. Para isso, dividiu-se a base de dados em dois conjuntos, o de treinamento e o teste, que foram sorteados de forma aleatória com semente fixa em 2104, representando 80% e 20% do total, respectivamente. O sorteio foi realizado considerando o atributo categórico que ainda não tinha sido utilizado, proximidade do oceano, para que houvesse proporção semelhante nos dois conjuntos.

Além da forma mais simples, também foi realizado o método k-fold. Esse método consiste na divisão da base de dados em K subconjuntos, com o intuito de fazer os testes um a um. No entanto, há uma desvantagem no valor escolhido para o K, pois um valor pequeno tende ao enveasamento do modelo, enquanto que um K grande tende a ter variabilidade nas medidas de desempenho. Assim, é recomendado usar o valor para K igual a 10, o qual foi o adotado nesse relatório.

Dessa forma, os modelos de regressão foram aplicados no conjunto de treinamento. Após isso, o conjunto de teste que estava reservado foi utilizado depois dos modelos prontos para realização das medidas de avaliação e predições. Foram avaliados o erro médio quadrático da raiz (RMSE), o médio absoluto (MAE) e o R-quadrado, este apresenta o quanto a variável dependente é explicada pelas independentes. Avalia-se, então, que o melhor modelo possui um menor RMSE e MAE somado a um maior R-quadrado.

4 Resultados e discussão

Iniciando com a divisão dos conjuntos de treinamento e de teste, o sorteio realizado ficou da seguinte forma:

Tabela 3: Conjuntos treinamento e teste

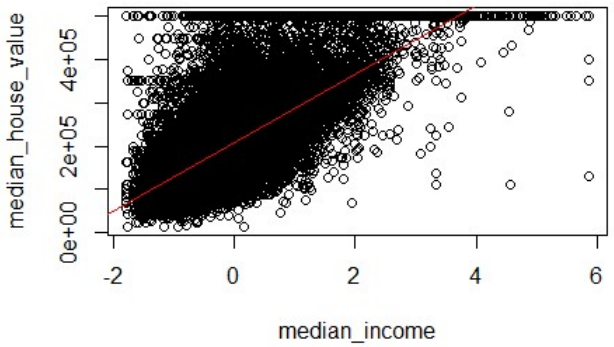
	<1H OCEAN	INLAND	ISLAND	NEAR BAY	NEAR OCEAN
Treinamento	7146	5215	4	1803	2067
Teste	1827	1310	1	458	531

4.1 Regressão Linear Simples

Para fazer a regressão linear simples foi utilizada a variável median_house_value como variável resposta, e median_income como variável explicativa. Inicialmente foi feito o gráfico de dispersão e pôde-se observar que há relação entre as variáveis com uma correlação positiva média, ou seja, quando a variável resposta aumenta, a variável explicativa também tende a subir. Afim de confirmar a análise, aplicou-se o teste de correlação, sendo

de aproximadamente 0.7. Com isso, é inferido que há correlação entre as duas, porém ela não é forte.

Figura 5: Gráfico de Dispersão



Seguindo com a análise, o modelo foi ajustado, de forma que ao analisar o resumo (summary), os testes T e F não rejeitaram a hipotese nula, provando que existe relação linear entre as váriaveis. Por sua vez, através da tabela da ANOVA, pôde-se observar que o parâmetro (beta 1) da variável median_income está bem estimado para a reta, pois através do p-valor rejeita-se a hipótese nula. Dessa forma, é possível confirmar a dependência linear entre o valor médio das casas e a renda média.

4.2 Regressão Múltipla

Para a regressão múltipla foram considerados todos os atributos, exceto a variável categórica. Dessa forma, analisando o modelo gerado, percebeu-se que o Teste T rejeitou a hipótese de que todos os coeficiente são diferentes de zero. Além disso, segundo o teste F, tiveram-se evidências suficientes para afirmar que existe pelo menos uma variável que mantém relação linear com a variável resposta. Ademais, o R-quadrado ajustado apresentou um bom resultado, no qual apresentou valor de aproximadamente 0,6356, mostrando que cerca de 63,56% da variabilidade total está sendo explicada pelo modelo de regressão.

Verificando a Normalidade

Com o modelo ajustado, foi feito a análise residual. Primeiramente, para verificar a normalidade foi realizado o teste de Lilliefors (Kolmogorov-Smirnov). Considerando o nível de significância de 5%, rejeitou-se a hipotese nula. Ou seja, não se têm evidências suficientes para afirmar que os dados seguem normalidade. Porém, através do gráfico pode-se constatar que os dados seguem certa normalidade, mas estes estão sendo afetados por possíveis outliers. Com base nessas duas afirmações, foi considerado a não normalidade dos resíduos.

Verificando a Homocedasticidade

Após a análise de não normalidade verificada, foi testado a homocedasticidade. Considerando um nível de confiança de 5%, ao realizar os teste de Breusch-Pagan, rejeitou-se

a hipótese nula. Ou seja, têm-se evidências suficientes para afirmar que os erros não são homocedásticos. Pelo gráfico é possível observar que os pontos não estão distribuídos de forma aleatória, comprovando, assim, a afirmação do teste.

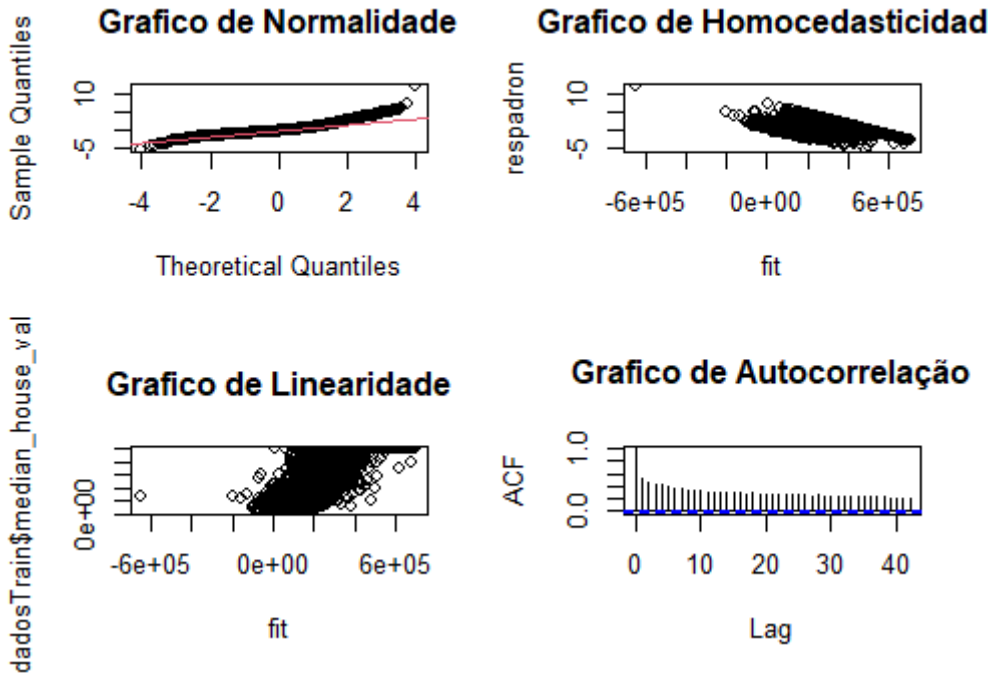
Verificando a Linearidade

Outro ponto da análise residual foi a verificação da linearidade. Ao analisar o teste RESET, percebemos que a hipótese de linearidade é rejeitada a um nível de significância de 5%. Além disso, ao ver o gráfico, pode-se considerar que a hipótese de linearidade está sendo violada, visto que os pontos distoam da reta.

Verificando a Autocorrelação

A última análise residual do modelo foi a verificação de autocorrelação dos erros. Por meio do teste Durbin-Watson, pôde-se observar que todas as correlações deram valores muito baixos, bem próximos a 0. De acordo com o gráfico, verifica-se que todos os lags estão fora dos limites, ou seja, a um nível de 5% de significância, rejeita-se a hipotese nula, implicando dizer que há evidências suficientes para afirmar que a suposição de independência dos erros está sendo violada.

Figura 6: Gráficos da Análise de Resíduos



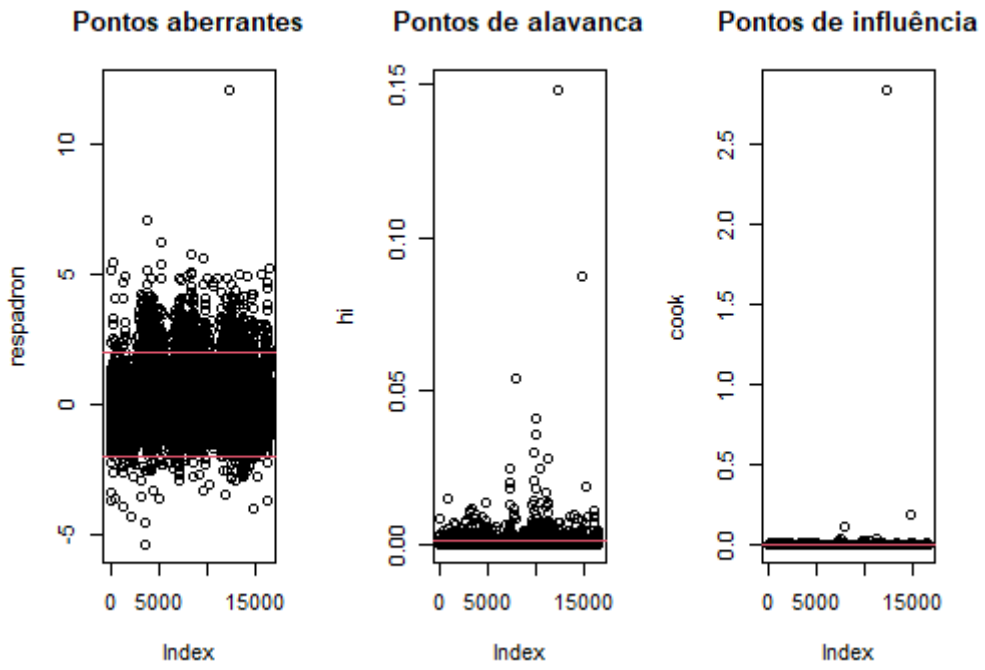
Retirando pontos

Após todo o processo de análise de resíduos, foi feita a análise de diagnóstico. Dado que o banco trabalhado há muitas observações, foi inviável identificar os pontos manualmente. Dessa forma, foram construídos códigos em R com o objetivo de identificar os pontos aberrantes, de alavanca e de influência. Para fazer uma análise mais aprofundada do

modelo, retirou-se todos os pontos aberrantes que não são de alavanca nem de influência. Assim feito, foi observado que há 338 observações que poderiam ser retiradas.

Sendo assim, um novo modelo sem os pontos identificados foi analisado. Percebeu-se que nos Testes T, continuou rejeitando a hipótese de que todos os coeficiente são diferentes de zero. Ademais, o R-quadrado ajustado apresentou um bom resultado e foi maior que o modelo anterior, no qual apresentou valor de aproximadamente 0,6479, o que mostra que cerca de 64,79% da varabilidade total está sendo explicada pelo modelo de regressâo. Mostrando uma pequena melhora do R-quadrado, somado a uma diminuição do sigma chapéu, pode-se concluir que esse novo modelo com os pontos aberrantes retirados apresenta uma pequena melhora de ajuste.

Figura 7: Gráfico dos pontos



Vale salientar que esses modelos implementados, por não passarem em nenhum dos pressupostos acabam se tornando inviáveis para as predições. Sendo assim, faremos uma nova seleção de variável através da multicolinearidade visando a construção de um modelo melhor.

Multicolinearidade

Multicolinearidade em regressão é uma condição que ocorre quando algumas variáveis preditoras no modelo estão correlacionadas a outras variáveis preditoras. A multicolinearidade forte é problemática porque pode aumentar a variância dos coeficientes de regressão, tornando-os instáveis. Sendo assim, foi feito o uso do VIF (fatores de inflação da variância), no qual medem o quanto a variância de um coeficiente de regressão estimado aumenta se seus preditores estão correlacionados, ou seja, o VIF mede a correlação da variável com todas as outras do modelo, sendo indicado eliminar aquelas com VIF>2.

Sendo assim, o uso da multicolinearidade tornou-se necessário para aplicar uma nova seleção de variáveis com o objetivo de melhorar o modelo já implementado.

Logo, após a seleção através do VIF, levou-se em consideração as seguintes variáveis: valor médio das casas, idade das casas e renda média. Pôde-se concluir que ainda há uma má implementação do model devido à queda do R-quadrado do modelo. Por outro lado, observa-se uma pequena queda do sigma chapéu, algo que é positivo para regressão. Sendo assim, visando novamente a melhora do modelo será aplicado uma tranformação da variável resposta, tendo como foco uma maior diminuição do sigma chapéu e o aumento do R-quadrado. Logo, será feito o uso do Box-Cox para a transformação citada.

Box Cox

A transformação Box-Cox é uma das possíveis formas de contornar o problema de dados que não obedecem os pressupostos da análise de capacidade, como por exemplo, normalidade dos dados. Esta transformação é dada por:

Figura 8: Transformação Box Cox

$$Y_i(\lambda) = \begin{cases} \ln(X_i), & \text{se } \lambda = 0, \\ \dfrac{X_i^{\lambda} - 1}{\lambda}, & \text{se } \lambda \neq 0, \end{cases}$$

Logo, através dessa fórmula, obtém-se a transformação da variável resposta, onde pode-se concluir que o modelo teve uma melhora significativa comparado aos anteriores, devido apresentar uma grande diminuição do sigma chapéu do modelo, apesar de também apresentar uma queda no R-quadrado. Pode-se provar isso com os gráficos dos pressupostos abaixo.

Verificação da Normalidade

Ao realizar o teste de lilliefors e considerando o nível de significância de 5%, rejeita-se a hipótese nula, ou seja, não se há evidências suficientes para afirmar que os dados seguem normalidade. Porém, sabe-se que quando trabalhado com muitas observações, os testes podem ser afetados pelos outliers, visto que os testes são muito sensíveis. Com isso, através do gráfico, pode-se constatar que os dados seguem normalidade. Com base nessas duas afirmações, e levando em conta a sensibilidade do teste, será considerado a normalidade dos resíduos.

Verificação da Homocedasticidade

Considerando um nível de confiança de 5%, ao realizar o teste de Breusch-Pagan, rejeita-se a hipótese nula, ou seja, há evidências suficientes para afirmar que os erros não são homocedásticos. Pelo gráfico é possível observar que os dados não se apresentam distribuídos de forma aleatória, comprovando, assim, a heterocedasticidade.

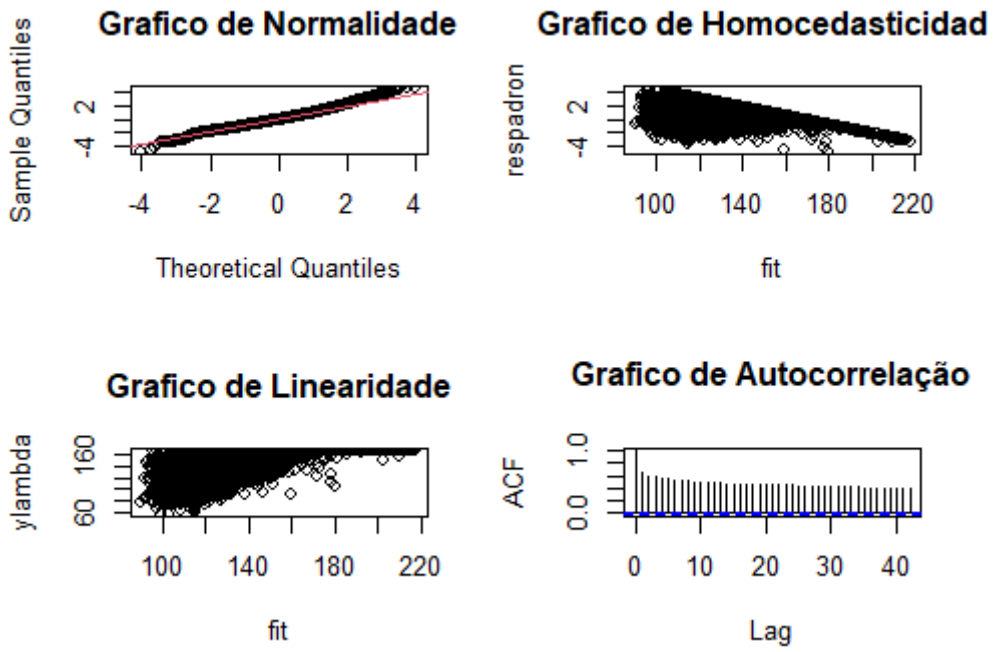
Verificação da Linearidade

Ao analisar o teste RESET, percebe-se que o p-valor deu menor que o nível de significância considerado de 5%, ou seja, a hipótese de linearidade é rejeitada. Além disso, ao ver o gráfico, é possível considerar que a hipótese de linearidade está sendo violada, visto que os pontos distoam de um reta.

Verificação da Autocorrelação

Pode-se observar por meio do teste de Durbin-Watson que todas as correlaçoes são muito baixas, proximas a 0, mostrando assim que a autocorrelação dos resíduos é muito baixa. De acordo com o gráfico, pode-se ver que todos os lags estão fora dos limites, ou seja, a um nível de 5% de significância, rejeita-se a hipótese nula, implicando dizer que há evidências suficientes para afirmar que a suposição de independência dos erros está sendo violada.

Figura 9: Gráfico da análise de resíduos



Validação Cruzada

Depois de todos os modelos concluídos, foi realizado a validação cruzada de duas formas afim de identificar qual o melhor modelo, o que está mais bem ajustado e com boas predições. Dessa forma, foram avaliadas as medidas do R-quadrado, RMSE e MAE para os principais modelos obtidos: MRLS e MRLM, sendo o primeiro ajuste com todos os atributos numéricos, o segundo com a seleção de variáveis feita pela multicolinariedade e por fim, com o boxcox. Destacando que os dois últimos tiveram a retirada de alguns outliers.

Com o método simplificado, é observado primeiramente, na tabela 4, que o modelo que possuiu o maior R-quadrado foi o primeiro ajuste da regressão linear múltipla. No

entanto, os erros também apresentaram um valor muito alto, o que é muito ruim para o modelo. Já o modelo realizado após a transformação de boxcox, é observado que apesar do R-quadrado ser menor comparado ao citado anteriormente, os erros se destacaram por serem os mais baixos de todos os modelos, mostrando, assim, ser melhor.

Tabela 4: Validação cruzada - método simplificado

R-quadrado	RMSE	MAE	Modelo
0,4770629	82582,82	61797,67	MRLS
0,6368851	68813,88	50165,35	MRLM
0,509299	80030,74	59544,75	MRLM com multicolinariedade
0,5092593	236337,1	206924	MRLM com boxcox

Nada muito diferente foi visualizado no método de k-fold. Os resultados dos erros e do R-quadrado apresentados na tabela 4 foram muitos semelhantes ao métodos mais simples da validação cruzada. O modelo de regressão linear múltipla após a transformação de boxcox ainda se destaca por ter erros muito baixos, indicando melhores previsões, e com relação ao R-quadrado, esse modelo continuou sendo mais preciso que a regressão linear simples.

Tabela 5: Validação cruzada - método k-fold

R-quadrado	RMSE	MAE	Modelo
0,472938	84015,19	62883,02	MRLS
0,6337874	70022,15	51180,54	MRLM
0,5194356	78061,2	58090,89	MRLM com multicolinariedade
0,4905648	15,07849	11,69102	MRLM com boxcox

5 Conclusão

Do trabalho desenvolvido, considerando o banco de dados dos valores das casas na Califórnia e com base os resultados obtidos, pode-se dizer que o modelo de regressão linear simples não se mostrou razoável, visto que a correlação entre as variáveis não foi alta. No modelo de regressão múltipla observou-se que nenhum pressuposto foi obedecido. Com isso, esse modelo também foi considerado inviável. A última alternativa foi o modelo considerando a multicolinearidade e o box-cox. Por meio dos resultados da análise residual desse modelo, pôde-se ver uma melhora significativa, levando a considerar como o melhor modelo para fazer predição.

Para confirmar a análise acima foi feito a validação cruzada. Por meio dela, pode-se concluir que o modelo com menores erros para fazer predição realmente é o último, modelo em que foi utilizado a multicolinearidade, assim como o box cox.

Referências

[1] California Housing Prices. Disponível em: *https://www.kaggle.com/camnugent/california-housing-prices* (Dezembro, 2020).

[2] Introduction to machine learning in R. NUGENT, C. Disponível em: *https://www.kaggle.com/camnugent/introduction-to-machine-learning-in-r-tutorial* (Dezembro, 2020).

[3] Avaliação de Modelos ou Classificadores. NETO, E. Disponível em: *http://www.de.ufpb.br/ eufrasio/Reg2/Aula_AvalModelos.pdf* (Dezembro, 2020).

[4] Aprendizagem de Máquina. FILHO, T. Disponível em: notas de aula (Dezembro, 2020).

[5] Validação cruzada em programação R. MISRH, R. Disponível em: *https://www.geeksforgeeks.org/cross-validation-in-r-programming/* (Dezembro, 2020).

[6] Seleção de Variáveis em Modelos de Regressão. CARVALHO, A. GOÉS, G. Disponível em: *https://repositorio.enap.gov.br/* (Dezembro, 2020).