UNIVERSIDADE FEDERAL DO CEARÁ
FACULDADE DE ECONOMIA, ADMINISTRAÇÃO, ATUÁRIA E CONTABILIDADE
CURSO DE GRADUAÇÃO EM CIÊNCIAS ECONÔMICAS

LUIZ CARLOS DE OLIVEIRA FILHO

DETERMINAÇÃO DE MODELO DE APRENDIZAGEM DE MÁQUINA PARA PRECIFICAÇÃO DE IMÓVEIS

FORTALEZA
2023

LUIZ CARLOS DE OLIVEIRA FILHO

DETERMINAÇÃO DE MODELO DE APRENDIZAGEM DE MÁQUINA PARA PRECIFICAÇÃO DE IMÓVEIS

Trabalho de Conclusão de Curso apresentado ao curso de Graduação em Ciências Econômicas, da UNIVERSIDADE FEDERAL DO CEARÁ, como requisito parcial para a Obtenção do grau de Bacharel em Economia.

Orientador: Prof. Dr. Sérgio Aquino de Souza

Fortaleza

2023

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

O48d Oliveira Filho, Luiz Carlos de.
Determinação de modelo de aprendizagem de máquina para precificação de imóveis / Luiz Carlos de Oliveira Filho. – 2023.
28 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Faculdade de Economia, Administração, Atuária e Contabilidade, Curso de Administração, Fortaleza, 2023.
Orientação: Prof. Dr. Sergio Aquino de Souza.

1. Pricing. 2. Properties. 3. Machine learning. I. Título.

CDD 658

LUIZ CARLOS DE OLIVEIRA FILHO

DETERMINAÇÃO DE MODELO DE APRENDIZAGEM DE MÁQUINA PARA PRECIFICAÇÃO DE IMÓVEIS

Trabalho de Conclusão de Curso apresentado ao curso de Graduação em Ciências Econômicas, da UNIVERSIDADE FEDERAL DO CEARÁ, como requisito parcial para a Obtenção do grau de Bacharel em Economia.

Fortaleza, __ de _____ de ____

BANCA EXAMINADORA

_____

Prof. Dr. Sérgio Aquino de Sousa

Universidade Federal do Ceará (CAEN/UFC)

_____

Prof. Dr. Rafael Barros Barbosa

Universidade Federal do Ceará (DEA/UFC)

_____

Prof. Dr. Ricardo Brito Soares

Universidade Federal do Ceará (CAEN/UFC)

Dedico este trabalho a Deus, minha família e aos professores que me orientaram.

AGRADECIMENTOS

Agradeço, como na dedicatória, principalmente a Deus, que apesar de mim, tem me abençoado sempre.

À minha família, em especial à minha esposa, que abdicou de várias horas de descanso para me apoiar durante a graduação.

Agradeço aos meus pais, que formaram meu caráter e meu intelecto, tornando possível este momento.

Agradeço também ao meu orientador, Prof. Dr. Sérgio Aquino de Souza, que para além da orientação, me possibilitou aprender tanto em tão pouco tempo. Também pelas ideias e apoio.

Agradeço ainda ao Prof. Dr. Glauber Nojosa, que tanto me ajudou em questões pertinentes à graduação, além de ter sido um excelente professor em minha formação.

Agradeço ainda ao amigo Fillipe Alencar, pelos conselhos e companheirismo, fundamentais nesse processo.

"E eles disseram: Crê no Senhor Jesus Cristo e serás salvo, tu e a tua casa."

(Atos dos apóstolos 16:31)

RESUMO

O uso de técnicas de aprendizado de máquina é oportuno para auxiliar na precificação de imóveis. Dentre os vários tipos de ferramentas do Machine Learning, faz-se necessário determinar qual responde melhor às necessidades da previsão de preços para o ramo imobiliário. Este, por sua vez, tem notável importância na economia brasileira. Entretanto, ainda são praticados métodos subjetivos e antiquados nessa atividade. O presente se propõe a definir o modelo de Machine Learning, entre os estudados, que melhor preveja preços para imóveis. Após submeter um modelo onde o preço é definido através de variáveis que caracterizam os imóveis, concluiu-se que a Regressão Lasso foi a que obteve melhor desempenho, apresentando os menores erros durante a projeção de valores de preço.

Palavras-chave: Precificação; imóveis; Machine learning.

ABSTRACT

The use of machine learning techniques is indeed opportune to assist in real estate pricing. Among the various types of machine learning tools, it is necessary to determine which one best meets the needs of determining prices for the real estate sector, which holds notable importance in the Brazilian economy. However, subjective and outdated methods are still practiced in this field. This study aims to define the machine learning model, among those studied, that best predicts property prices. After applying a model where the price is defined by variables representing the aspects considered during negotiations, it was concluded that Lasso Regression performed with the lowest errors in projecting price values.

Keywords: Pricing; Properties; Machine Learning

LISTA DE ILUSTRAÇÕES

LISTA DE ABREVIATURAS E SIGLAS

MQO	Mínimos Quadrados Ordinários
REQM	Root-Mean-Squared Error
EQM	Mean-Squared Error
ABNT	Associação Brasileira de Normas Técnicas
NBR-14653:2001	Norma Brasileira 14653:2001
IPTU	Imposto Predial e Territorial Urbano

SUMÁRIO

1 **INTRODUÇÃO**

No Brasil, a aquisição de imóveis é um processo delicado, por parte do consumidor, devido ao fato de que este bem consome parte considerável da renda média. Este fato fica mais evidente se compararmos o salário mínimo no país e o valor médio de uma residência de baixo valor. Se por um lado o consumidor busca seus interesses, o ofertante não deixa de fazê-lo: Entre brasileiros, a aquisição de imóveis para fins comerciais é uma das opções preferidas de investimento, seja vendendo, alugando ou mesmo ter como reserva de valor (ALENCAR, 2022).

O conflito de interesses durante a negociação enfrenta ainda questões mais difíceis devido às diferentes características que podem agregar valor à propriedade, podendo as características serem de ordem quantitativa ou qualitativa. É sabido que a Associação Brasileira de Normas Técnicas (ABNT) determina um padrão para estabelecimento de valor de mercado para esse tipo de construção, através da norma NBR-14653:2001. Entretanto, como mencionado, aspectos não numéricos tem claras influências sobre a ideia de valor que o lugar transmite. Para todos os efeitos, mensurações equivocadas para valores de residência causam perda de bem-estar para ambas as partes: consumidores, que não conseguem adquirir a residência e para os ofertantes, que não conseguem vendê-las. A situação pode agravar ainda o déficit habitacional vivido no território brasileiro, porém a este último, o presente trabalho não irá se deter.

Dadas as problemáticas apresentadas faz-se necessário um método capaz de prever preços para imóveis de forma objetiva, ágil e que englobe características procuradas pelo consumidor, como número de quartos, banheiros, etc. Este método, por sua vez, precisa também dispor de capacidade de ser aprimorado em relação à dinâmica do mercado, ou seja, ajustar-se com os erros. Para Shihao Gu, et al (2020), o aprendizado de máquina propõe modelos de alta dimensionalidade para previsão estatística, métodos de "regularização" para seleção do modelo e mitigação do ajuste excessivo, além de algoritmos que selecionam especificações de modelos potenciais (2020). Desse modo, o presente trabalho utiliza métodos de Machine Learning para prever preços e seleciona a regressão que melhor desempenha essa função para o contexto específico da precificação para o mercado imobiliário. Para tanto, são testadas as previsões de preço por meio de regressões em MQO, Ridge e Lasso.

1.1 OBJETIVOS

Não sendo as abordagens tradicionais suficientes para elaborar uma abordagem precisa e equilibrada para a precificação do setor imobiliário (a saber, pesquisas de mercado com observações limitadas, médias de preço e atribuições subjetivas dos proprietários e anunciantes), o presente trabalho almeja propor, através de uma abordagem orientada a dados, uma metodologia de precificação mais adequada e objetiva.

O objetivo geral, por sua vez, é definir a regressão que melhor prevê preços para imóveis, dentro do modelo abordado. Como objetivos específicos podemos elencar:

- Previsão de preços através do método de Mínimos Quadrados Ordinários
- Previsão de preços através da regressão Ridge
- Previsão de preços através da regressão Lasso
- Definição das características que melhor explicam a variável alvo
- Avaliação do EQM dos métodos anteriores
- Análise e escolha dos melhores resultados entre os métodos elencados

2 **MÉTODOS EMPÍRICOS**

Nesta seção, dispõem-se os conceitos pertinentes à compreensão e elaboração do presente trabalho. Inicialmente, é discutido o contexto do tema no mercado brasileiro e, em seguida, aborda-se o conceito de aprendizado de máquina propriamente dito. São incluídos também preceitos sobre as ferramentas utilizadas. Vale ressaltar que, para fins didáticos, a abordagem do assunto Machine Learning não será exaustiva, abrangendo apenas os tópicos de interesse do estudo.

2.1 MÉTODOS DE APRENDIZAGEM

Com o crescente volume de dados produzidos mundialmente, a inteligência artificial tem atingido cada vez mais relevância. Para fomentar o uso dessas tecnologias, foram desenvolvidas ferramentas que permitem identificar padrões para projetar resultados futuros, no que passamos a chamar de aprendizado de máquina. Esse aprendizado utiliza-se de ferramentas estatísticas e, dentre estas estão as regressões, que são capazes de, entre outras funcionalidades, selecionar as variáveis relevantes, ajustar erros de previsão, identificar padrões, etc.

2.2 PRECIFICAÇÃO DE IMÓVEIS NO BRASIL

No país, a despeito de efeitos adversos como a pandemia causada pelo SARS-COV e crises que atingiram o mercado imobiliário direta ou indiretamente, o setor cresceu de forma notável nos últimos 4 anos. Tal resultado é consistente a ponto de haverem projeções de permanência de crescimento para os próximos anos.

Segundo Araruna (2022), estes resultados no ramo são apoiados, entre outros motivos pelas políticas de expansão de crédito em anos recentes, o valor da taxa Selic e, além dos fatores financeiros, as medidas de isolamento social, que afetaram as expectativas do consumidor em relação à aquisição de imóveis. O aumento na demanda, por sua vez, impulsionou a procura por mão-de-obra qualificada, que não sendo suficiente, permitiu que muitos corretores inexperientes fossem inseridos no

mercado. Este último tem tornado distorcidas as médias de preço e os valores propostos para venda e locação.

2.2.1 **Regressão Linear Múltipla e método dos Mínimos Quadrados Ordinários**

O modelo de regressão linear múltipla permite estimar a relação entre uma variável dependente e duas ou mais variáveis independentes. A ideia por trás desse modelo estatístico é que, mantendo-se outras variáveis fixas (*ceteris paribus*), podemos estimar o efeito de cada variável independente sobre a variável dependente (Chein, 2019). A regressão linear múltipla pode ser aplicada em diversos campos de estudo, sobretudo econométricos, como a avaliação do efeito da temperatura no verão sobre o PIB na agropecuária ou o efeito dos anos de estudo sobre o trabalho. No caso particular deste trabalho, por exemplo, pôde-se observar os efeitos de variáveis, como número de quartos, sobre o preço de imóveis de uma determinada cidade. Não à toa, O modelo mais usado quando se desejar calcular o comportamento de uma variável dependente é o de regressão linear (Silva, Gustavo 2019).

A regressão linear visa compreender a relação (ou não) entre uma variável dependente Y e suas variáveis independentes. Pode ser algebricamente expressa por:

$$Y_i = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 + \cdots X_{ip}\beta_p + \epsilon_i \qquad (1)$$

onde x_1 , x_2 , ..., x_p são as variáveis independentes utilizadas para prever Y_i; β_0 é o intercepto, ou seja, o valor de Y quando todas as variáveis independentes são iguais a zero. β_1 , β_2 , ..., β_p são os coeficientes, que representam a relação entre as variáveis independentes e dependente. Por fim, ε_i é o termo de erro, que engloba o efeito de fatores não explicados pelas variáveis independentes.

2.2.2 **Mínimos Quadrados Ordinários**

O método dos mínimos quadrados ordinários nada mais é do que uma ferramenta que minimiza a soma dos quadrados dos resíduos, ou seja, reduz a soma dos quadrados das diferenças entre os valores que um determinado modelo estima e

os valores observados. Este efeito permite encontrar uma aproximação linear entre a variável dependente e uma ou mais variáveis independentes.

O MQO seleciona parâmetros de modo que a soma dos quadrados dos resíduos seja a menor possível em qualquer modelo. Dessa forma, o método dos Mínimos Quadrados proporciona as estimativas dos parâmetros que fornecem o menor valor possível para a soma dos quadrados. Aos valores desses parâmetros, chamamos estimadores. Em outras palavras, o método minimiza o comprimento do vetor de erros (DOS SANTOS, 2017).

> Com a regressão é possível estimar o grau de associação entre Y, variável dependente e Xi, conjunto de variáveis independentes (explicativas). O objetivo é resumir a correlação entre Xi e Y em termos da direção (positiva ou negativa) e magnitude (fraca ou forte) dessa associação. Mais especificamente, é possível utilizar as variáveis independentes para predizer os valores da variável dependente. Em regressões multivariadas – compostas de mais de uma variável independente – é possível também identificar a contribuição de cada variável independente sobre a capacidade preditiva do modelo como um todo. Tecnicamente, dizer que o modelo é ajustado utilizando a forma funcional de mínimos quadrados ordinários significa que uma reta que minimiza a soma dos quadrados dos resíduos será utilizada para resumir a relação linear entre Y e X (FIGUEIREDO *et al.*, 2011).

A soma quadrática dos resíduos é:

$$\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2 \qquad (2)$$

O objetivo de MQO é encontrar os valores de β_j (*j=0,....,p*) da equação 1, que minimizam a soma da equação 2, resultando em uma linha de regressão que se ajusta melhor aos dados.

O MQO, propriamente dito, pode ser expresso como a solução do seguinte problema de minimização:

$$Min_{\beta} \sum_{i=1}^{n}\left(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j \; X_{ij} \right) \qquad (3)$$

2.2.3 **Regressão Lasso**

O *least absolute shrinkage and selection operator* (LASSO) é um modelo de regressão que foi proposto, entre outros objetivos, para solucionar problemas relacionados à alta dimensionalidade, ou seja, a condição de possuir um número de covariáveis muito elevado, possivelmente superior ao número de observações. Sendo um dos instrumentos estatísticos mais utilizados para este fim, o Lasso adiciona restrição na equação de mínimos quadrados, fazendo com que os coeficientes das variáveis dependentes tendam a zero à medida que aumenta o grau de penalização dos coeficientes. Este método estima e seleciona as covariáveis mais adequadas para o modelo (ALCÂNTARA, 2021).

O modelo pode ser expresso por:

$$Min_{\beta} \sum_{i=1}^{n} (Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j \, X_{ij}) \; + \; \lambda \sum_{i=1}^{p} |\beta_j| \qquad (4)$$

Onde o (hiper) parâmetro λ controla o grau de penalização: quanto maior, menores serão os valores dos coeficientes, podendo zerar alguns $\beta_{j's}$, o que resulta na exclusão da variável *j* do modelo. O modelo Lasso, portanto, encolhe coeficientes e seleciona variáveis.

2.2.4 **Regressão Ridge**

Tal como o modelo Lasso, uma das grandes vantagens do Ridge é a possibilidade de tratamento de dados. A regressão Ridge é um recurso estatístico importante para a análise de dados, por ser capaz de tratar problemas de multicolinearidade (MASCHKE *et al.*, 2018). Para tanto, a regressão Ridge também adiciona penalização ao estimador dos coeficientes de regressão encontrados pela minimização dos quadrados (ACOSTA; AMOROSO, 2021). No entanto, penalização ocorre na forma quadrática.

$$Min_{\beta} \sum_{i=1}^{n} (Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j \, X_{ij}) \; + \; \lambda \sum_{i=1}^{p} \beta_j^2 \qquad (5)$$

2.2.5 **Avaliação da precisão do modelo**

No processo de aprendizado de máquina, várias informações são analisadas a fim de buscar padrões, elaborar previsões e testar as mesmas. É intrínseco e necessário ao processo o aparecimento e ajuste de erros. Entretanto, podemos observar maior incidência de erros em certas formas de regressão em relação a outras. Para analisar a qualidade de um modelo de regressão é necessário, entre outras tantas medidas, avaliar os erros de ajuste. De fato, os modelos de previsão elaboram previsões de valores para a variável alvo através de aprendizado, apesar de que as estimativas produzirão valores que podem divergir em relação ao esperado. Tal comportamento, torna é imprescindível a checagem da aderência de cada modelo (SILVA, 2019).

Segundo Wang e Lu (2018), para avaliação, o Erro Quadrático Médio (EQM) é amplamente adotado em muitos sistemas, além da vantagem de ser capaz de medir a diferença entre as pontuações previstas e as avaliações reais dos usuários.

O EQM objetiva medir a média do quadrado dos resíduos, observando o quanto as predições se desviam, em média, dos valores observados. A EQM pode ser matematicamente traduzido por:

$$\frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \widehat{Y}_i\right|^2 \qquad (6)$$

Podendo ainda ser expresso, simplesmente, pela soma dos resíduos ao quadrado dividida pelo número de observações, ou seja:

$$\frac{\sum\left(Y_i - \widehat{Y}_i\right)^2}{n} \qquad (7)$$

Onde n é o número total de observações, y representa os valores reais da variável alvo e ŷ, os valores previstos para a variável dependente.

A raiz do erro quadrático (REQM) é outra das muitas métricas usadas para avaliar os erros de um modelo de regressão. Ele mede a raiz quadrada da média dos erros quadrados entre as previsões do modelo e os valores reais da variável alvo. Uma das grandes vantagens desta ferramenta é a de penalizar a variância, ajustando o modelo, uma vez que é capaz de atribuir maior peso a erros cujos valores absolutos são maiores em relação aos demais. De igual modo, erros menores tem pesos menores atribuídos a si (CHAI; DRAXLER, 2014).

O REQM pode ser definido por:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2} \qquad (8)$$

onde, há uma amostra de n observações y (yi, i = 1,2,..., n) e n previsões correspondentes do modelo y.

Para ambos, tanto EQM como REQM, os resultados de previsão serão tanto melhores quanto seus resultados forem menores. Menores resultados significam que o modelo está produzindo menores resíduos. É possível ainda calcular as duas medidas para conjuntos de treino e teste.

Outra possibilidade é identificar um possível *overfitting,* que estaria ocorrendo caso o conjunto de testes apresente EQM muito elevado em relação ao calculado no conjunto de treino.

O EQM evidencia outliers nos erros de forma mais clara. Ora, os valores são elevados ao quadrado, evidenciando discrepâncias em torno da média. Além disso, o erro quadrado médio pode ser utilizado no conjunto de testes e no conjunto de treinamento, apesar de que, para as análises pertinentes ao presente trabalho, são os valores no conjunto de teste que tem maior relevância (DE FARIAS, 2021).

3 **METODOLOGIA**

Para o presente estudo, foi extraída, de um site de compra e aluguéis de imóveis "Zap imóveis", uma base de dados contendo preços e características de diversas casas e apartamentos. A base contida está presente na plataforma Kaggle, no link https://www.kaggle.com/datasets/williamu32/imoveis-goiniago. Segundo Mattar (2005), esse processo configura uma obtenção de dados secundários A saber, as características apresentadas no catálogo do site são: endereço, área total, número de banheiros, número de quartos, valor do IPTU, número de vagas de garagem, o próprio preço, e o tipo de imóvel.

Foram definidos, então, modelos de regressão para realizar a previsão dos preços dos imóveis contidos nessa base e de dados. A escolha levou em consideração a frequência com que os modelos são mencionados e utilizados em produções semelhantes a esta. Deste modo, foram selecionados 3 modelos de regressão: regressão linear múltipla, ajustada por MQO; regressão Ridge e regressão Lasso. Estes são amplamente utilizados para precificação de imóveis.

Ao todo a amostra continha 13.031 observações, antes do tratamento de dados, que será descrito mais adiante. Também é importante considerar que a amostra contém dados do ano de 2021 e que, como o próprio trabalho sugere, a localidade das construções presentes observadas é o município de Goiânia, em Goiás. A pesquisa é dividida em x etapas: tratamento de dados e escolha de variáveis; divisão dos dados em bases de treino e teste; submissão do modelo ao método de Mínimos Quadrados Ordinários (MQO), conforme subseção 2.2.2; submissão do modelo à regressão RIDGE, conforme subseção 2.2.3; submissão do modelo à regressão LASSO, conforme subseção 2.2.4.

3.1 TRATAMENTO DE DADOS E ESCOLHA DE VARIÁVEIS

A primeira ação tomada durante o tratamento foi a remoção de valores vazios contidos na tabela obtida. Em um segundo momento, variáveis de pouca relevância foram removidas do modelo. A saber, as variáveis removidas foram a data de publicação, endereço e tipo de apartamento. A data de publicação do anúncio será de pouco interesse, uma vez que todas as observações são de 2021. Por razões

similares, o endereço das construções também foi removido, tendo em vista que as observações pertencem ao mesmo município. Por fim, dadas as premissas do estudo, apenas apartamentos foram observados, tornando desnecessária a inclusão da variável “tipo”, no modelo, dado que o único tipo que interessa é “apartamento”. Desse modo, as variáveis restantes serão analisadas no modelo, sendo elas: área total, número de banheiros, número de quartos, valor do IPTU, número de vagas de garagem e preço.

A variável “preço” teve seu formato alterado para valores numéricos. Quanto ao valores faltantes no dataframe, estes receberam imputações artificiais de valores e, através do método "Predictive Mean Matching" (pmm), valores médios foram previstos e inseridos no modelo.

3.2 DIVISÃO EM DADOS DE TREINO E TESTE

Feitos os devidos tratamentos da etapa anterior, a base de dados foi dividida em dois grupos: amostras de treino e teste. As amostras de treino foram construídas a partir de 70% do total de observações, ao passo que as amostras de teste receberam os demais 30% do total de observações.

3.3 TESTES ENTRE OS MODELOS

Para evitar não-linearidades e o não funcionamento de regressões que dependem da linearidade da equação como premissa, a variável ”preço” foi colocada em formato logarítmico. Dessa forma a regressão linear múltipla, entre outras, pode ser executada no modelo.

À exceção da variável “preço”, todas as demais foram submetidas a interações quadráticas entre si. O procedimento permite expandir o número de variáveis independentes, o que permite melhor funcionamento de regressões de penalização. Semelhante à etapa anterior e com o mesmo propósito, interações cruzando as próprias variáveis umas às outras também foram realizadas. Ao contrário das interações quadráticas, o resultado das interações entre as variáveis não é um quadrado de uma única variável, mas o produto de 2 variáveis diferentes, por exemplo "Areas:Bathrooms".

Em seguida, uma matriz de design foi criada com as variáveis explicativas e suas respectivas interações, enquanto a variável dependente foi mantida em formato logarítmico.

Em terceiro momento, as novas variáveis, obtidas pela operação de obtenção dos quadrados das variáveis originais, foram renomeadas com a inclusão do sufixo "_2". Assim, o quadrado da coluna "AREAS" passou a se chamar "AREAS_2", conforme quadro abaixo:

Quadro 1 — Variáveis quadráticas adicionadas ao modelo

VARIÁVEIS INICIAIS	VARIÁVEIS ADICIONADAS
Areas	Areas_2
Bathrooms	Bathrooms_2
Bedrooms	Bedrooms_2
Condominio	Condominio_2
IPTU	IPTU_2
Parking_Spaces	Parking_Spaces_2

Fonte: O autor (2023).

Noto total, após incluídas as interações e os termo quadráticos, o modelo possui 27 variáveis mais a constante. Foram então construídos modelos de regressão para prever valores para a variável alvo, analisar as relações entre as variáveis e analisar os erros de cada modelo em relação às suas previsões, a fim de entender qual deles melhor se ajusta para previsões do modelo. O primeiro, como já citado, é o d método de mínimos quadrados ordinários. As previsões para a variável resposta foram criadas, com base em um modelo que foi anteriormente ajustado, para o conjunto de teste.

As regressões de Ridge e Lasso foram submetidas a análises de parâmetro de regularização. Este (hiper) parâmetro foi obtido a partir de *cross-validation* e utilizado para ajustar as previsões dos valores da variável dependente. Foram verificadas as variáveis penalizadas ou não, bem como as restantes selecionadas por cada tipo de regressão.

4 RESULTADOS E DISCUSSÕES

Os testes entre as regressões e seus respectivos erros foram feitos através da linguagem R. Após os devidos tratamentos, a base de dados, que continha nove variáveis mais uma variável resposta, passou a compor sete, incluindo “preço”. Com a adição de interações entre estas seis variáveis remanescentes e os quadrados das mesmas, o modelo passou a ter 28 variáveis, estando mais robusto e suscetível a regressões que permitem penalizações.

Nesta seção, foram analisados os resultados dos modelos, valores ótimos do (hiper) parâmetro lambda (λ), variáveis selecionadas e penalizadas, bem como os resultado do EQM.

Foi inicialmente realizada uma regressão linear para analisar a relação entre as variáveis. Essa regressão foi ajustada através do método de Mínimos Quadrados Ordinários. Os coeficientes produzidos foram postos em um vetor “MQO.FIT.COEF” e uma coluna de “1”´s foi inserida na matriz de dados (variáveis independentes). As previsões para a variável resposta foram criadas, com base no modelo anteriormente ajustado, para o conjunto de teste.

Como a regressão MQO não seleciona variáveis, todas foram mantidas. Para este primeiro modelo, o valor do erro quadrado médio obtido na amostra teste obtido foi de 1,9183.

Previsões também foram criadas com o modelo de regressão Ridge. Para tanto, utilizou-se um valor de alpha igual a 0, no comando pacote *glmnet do R*. Para encontrar ótimo valor de λ, que regulariza o modelo, foi conduzido um *cross-validation*.

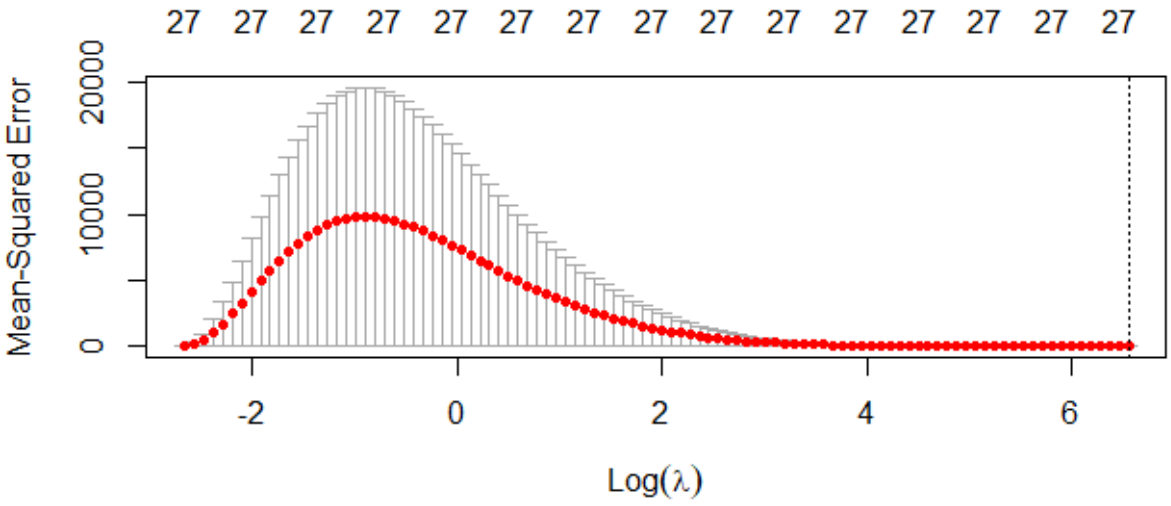
Dadas as características da regressão Ridge, esta não reduz o número de variáveis, mas sim as suas magnitudes, para chegar ao menor erro de ajuste. Em outras palavras, o modelo não seleciona variáveis. Como este modelo não seleciona variáveis, permanecem as 27 ao longo de toda a regressão .

À medida que lambda aumenta, o EQM também varia. O valor de λ selecionado é aquele que minimiza o EQM na amostra teste Lambda. Foi selecionado através do código desenvolvido em R. A saber, este valor corresponde ao situado na linha pontilhada vertical do gráfico na figura 1.

A fim de obter uma escala de melhor interpretação, foi utilizado o logaritmo de λ, ou Log λ. Desta forma a escala tem maior amplitude e o gráfico, melhor visualização. Os valores encontrados após a regressão foram os seguintes: para o λ, o modelo determinou o valor de 699.59, o que corresponde a Logλ= 6,55 e um EQM de 0.8628.

O valor do parâmetro λ foi utilizado para regularizar o modelo e realizar as previsões da variável alvo, com os menores resíduos possíveis. Obviamente, as variáveis são as mesmas iniciais, pois não houve seleção de variáveis, como mencionado anteriormente. O EQM=0.8628 se mostrou menor em relação à regressão anterior.

Figura 1 — Variáveis na regressão Ridge



Fonte: O autor (2023).

Adicionalmente , previsões também foram realizadas por meio do modelo de regressão Lasso. Para tanto, mudou-se o um valor de alpha para 1, no comando pacote *glmnet do R*. Para calcular valor ótimo de λ também foi conduzido um *cross-validation*.

Dadas as características específicas da regressão Lasso, o número de variáveis é tipicamente reduzido em relação ao conjunto inicial com 27 variáveis. Em outras palavras, o modelo seleciona variáveis.

À medida que lambda aumenta, o EQM também varia. O valor de λ selecionado é aquele que minimiza o EQM na amostra teste. Foi selecionado através

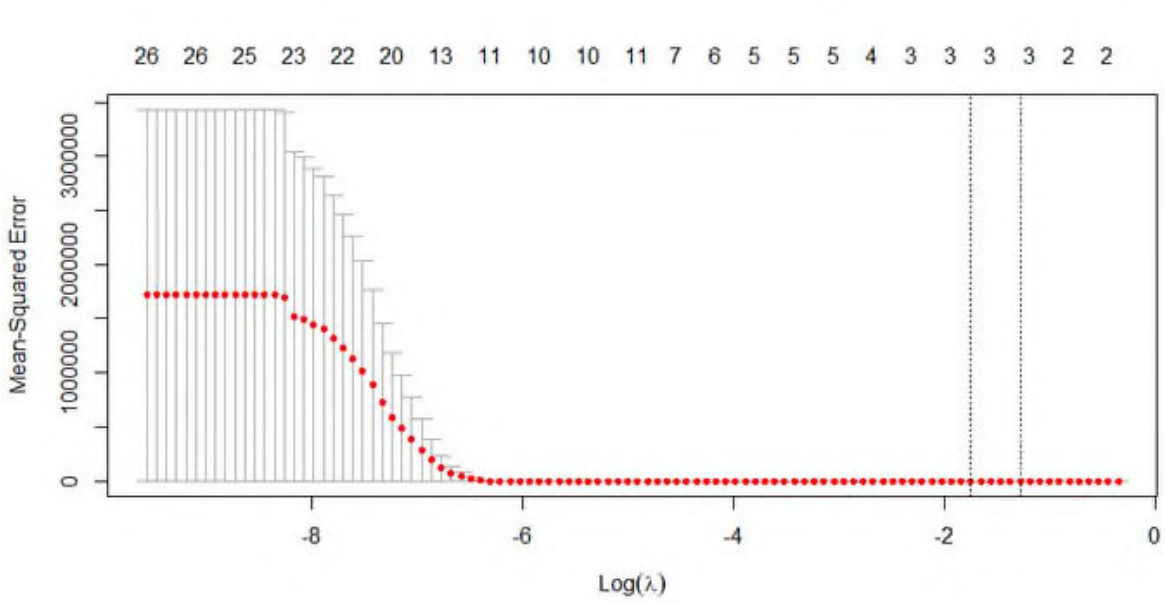
do código desenvolvido em R. A saber, este valor corresponde ao situado na linha pontilhada vertical mais à esquerda do gráfico na figura 2.

Foi obtido valor de λ=0,1733, ou seja, Log λ=-1.7528. Após seleção de variáveis, através da otimização penalizada, permaneceram no modelo a constante (que por definição não sofre penalização ou seleção) e duas variáveis: BATHROOMS e BEDROOMS.

A fim de obter uma escala de melhor interpretação, foi utilizado o logaritmo de λ. Desta forma a escala tem maior amplitude e o gráfico, melhor visualização.

O valor do parâmetro λ foi utilizado para otimizar o modelo e realizar as previsões da variável alvo, com os menores resíduos possíveis. Nota-se que, as variáveis não são as mesmas iniciais, pois houve seleção de variáveis, como mencionado anteriormente. Encontrou-se um EQM=0,6231, valor menor em relação aos dois modelos anteriores.

Figura 2 — Variáveis da regressão Lasso



Fonte: O autor (2023).

5 **CONCLUSÃO**

A regressão linear, apesar de ajustada com mínimos quadrados, reportou um EQM de 1,9183. Este valor, entre as regressões analisadas, foi o maior. O modelo também não selecionou variáveis, mantendo todas. Para esta etapa, apesar de não haver seleção de variáveis, 13 colunas foram identificadas como estatisticamente relevantes a um nível de significância de 5%

A regressão Ridge, por sua vez, também não selecionou variáveis, mantendo as vinte e sete. O valor de erro quadrático médio foi de 0,8628. Apesar de ser bem menor em relação ao obtido anteriormente, o EQM permanece elevado.

Já, com a regressão Lasso, houve interessante penalização de variáveis, mantendo apenas duas, das vinte e sete iniciais. A regressão, além da penalização, selecionou variáveis. Das vinte e sete variáveis presentes, o Lasso manteve apenas duas, além do intercepto. A saber, as variáveis mantidas foram: "banheiros", "quartos" e o próprio intercepto do modelo foram selecionados. A escolha está dentro do que se esperava observar, dado que as características estão presentes nos anúncios mais vistos e nas negociações do consumidor. Quanto aos erros retornados, o EQM apresentado pela regressão foi de 0,6231.

Quadro 2 — Comparativos dos erros REQM

Regressão linear múltipla com MQO	RIDGE	LASSO
1,9183	0,8628	0,6231

Fonte: O autor (2023).

Dadas as circunstâncias, conclui-se que, para este modelo de precificação de imóveis, entre as ferramentas analisadas (regressão linear múltipla, Ridge e Lasso), a regressão Lasso é a que possui melhor desempenho.

REFERÊNCIAS

ABBAD, Gardênia ; TORRES, Cláudio Vaz . Regressão múltipla stepwise e hierárquica em Psicologia Organizacional:: aplicações, problemas e soluções. **Estudos de Psicologia**, Natal, v. spe. 11 p, 10 set 2002.

ACOSTA, Simone; AMOROSO, Anderson. Aplicação da regressão por vetores de relevância na modelagem de um processo produtivo*. In:*
MARTINS, Ernane (Org.). **ENGENHARIA DE PRODUÇÃO**: PLANEJAMENTO E CONTROLE DA PRODUÇÃO EM FOCO. Guarujá: Editora Científica Digital, v. 1, 2021. 268 p. cap. 3, p. 40-54. Disponível em: https://downloads.editoracientifica.com.br/books/978-65-87196-71-8.pdf. Acesso em: 17 jun. 2023.

ALCÂNTARA, Gilberto. **Avaliação do lasso e métodos alternativos em modelos de regressão logística**. São Carlos, 2021. 121 p Dissertação (Mestrado em Estatística) - Universidade Federal de São Carlos, São Carlos, 2021.

ALENCAR, Sérgio Ricardo Ribeiro . **Precificação de Imóveis com Machine Learning**. São Carlos, 2022 Trabalho de Conclusão de Curso (MBA em Inteligência Artificial e Big Data) - Universidade de São Paulo, São Carlos, 2022.

ARARUNA, Rafael Santana . **Análise Imobiliária**: Qual o melhor método para prever o valor de um imóvel?. Brasília, 2022. 87 p Trabalho de Conclusão de Curso (Estatística) - Universidade de Brasília, Brasília, 2022.

CHAI, T. ; DRAXLER, R. R. . **Root mean square error (REQM) or mean absolute error (MAE)?** : Arguments against avoiding REQM in the literature. Arguments against avoiding REQM in the literature. Maryland, 2014. 4 p. Disponível em: https://gmd.copernicus.org/articles/7/1247/2014/. Acesso em: 8 jul. 2023.

CHEIN, Flávia . **Introdução aos modelos de regressão linear**: Um passo inicial para compreensão da econometria como uma ferramenta de avaliação de políticas públicas. Brasília: Enap, 2019. 76 p. (Metodologias de pesquisa).

DE FARIAS, DANIEL. **ANÁLISE DA CAPACIDADE PREDITIVA DE MODELOS DE REGULARIZAÇÃO APLICADO AO IPCA**. Fortaleza, 2021 Dissertação (Pós Graduação em Economia) - Universidade Federal do Ceará, Fortaleza, 2021.

DOS SANTOS, LEVI ALÃ NEVES . **MÍNIMOS QUADRADOS ORDINÁRIOS (MQO) NA PRODUÇÃO CIENTÍFICA BRASILEIRA:**: A INTERDISCIPLINARIDADE ENTRE A ECONOMETRIA E AS METRIAS DA INFORMAÇÃO (BIBLIOMETRIA,

INFORMETRIA E CIENTOMETRIA). Salvador, 2017. 189 p Tese (Pós Graduação em Ciência da Informação) - Universidade Federal da Bahia, Salvador, 2017.

FIGUEIREDO, Dalson *et al*. O que Fazer e o que Não Fazer com a Regressão: pressupostos e aplicações do modelo linear de Mínimos Quadrados Ordinários (MQO). **Revista Política Hoje**, v. 20. 99 p, 2011.

GU, Shihao ; KELLY, Bryan ; XIU, Dacheng . Empirical Asset Pricing via Machine Learning . **The Review of Financial Studies**, v. 33, n. 5. 50 p, 26 fev 2020.

HODSON, Timothy O. . **Root-mean-square error (REQM) or mean absolute error (MAE):** : when to use them or not. Copernicus.org. Illinois, 2022. 7 p. Disponível em: https://gmd.copernicus.org. Acesso em: 8 jul. 2023.

MASCHKE, Maurício de Conto *et al*. Fatores determinantes da estratégia de preços da soja através da Regressão Ridge*. In:* VI SIMPÓSIO DA CIÊNCIA DO AGRONEGÓCIO, n. 6. 2018. 2018. 10 p.

PASSOS, Luiz Fernando Coelho . **MÉTODOS DE REGULARIZAÇÃO NO APRENDIZADO DE MÁQUINAS: RIDGE E LASSO**. Niterói, 2022 Trabalho de Conclusão de Curso (Estatística) - Ime - Instituto de Matemática e Estatística, Niterói, 2022.

SILVA, Gustavo. **Modelos de aprendizagem de máquina para precificação de imóveis na cidade de Fortaleza**. Fortaleza, 2019 Trabalho de Conclusão de Curso (ENGENHARIA CIVIL) - Universidade Federal do Ceará, Fortaleza, 2019.

WANG, Weijie ; LU, Yanmin. Analysis of the Mean Absolute Error (MAE) and the Root Mean Square Error (REQM) in Assessing Rounding Model. **IOP Conference Series: Materials Science and Engineering**, Kuala lumpur, v. 324, 21 fev 2018. Disponível em: https://iopscience.iop.org/article/10.1088/1757-899X/324/1/012049. Acesso em: 8 jul. 2023.

APÊNDICE A — Script utilizado em linguagem R

```r
rm(list = ls())

library(mice)
library(glmnet)
dfgo<- read.csv("C:/Users/s/Downloads/base_tratada_goiania (2).csv", comment.char="#")

colSums(is.na(dfgo))

dfgo$PRICE[dfgo$PRICE == 'consulta'] = NA
dfgo$PRICE=as.numeric(dfgo$PRICE)
unique(dfgo$TIPO)

dfgo =subset(dfgo, select = -c(ADDRESS, DATE, TIPO))


imp <- mice(dfgo, seed = 123, method='pmm')

dfgo_imputed <- complete(imp)
colSums(is.na(dfgo_imputed))

dfgo_imputed$LOGPRICE=log(dfgo_imputed$PRICE)

dfgo_imputed =subset(dfgo_imputed, select = -PRICE)

```

```r

colnames(dfgo_imputed)
aux=LOGPRICE~ (AREAS+BATHROOMS+BEDROOMS + CONDOMINIO  +  IPTU+ PARKING_SPACES )^2
mat_interact=model.matrix(aux, data=dfgo_imputed)[,-1]

aux2= subset(dfgo_imputed^2, select = -LOGPRICE)

colnames(aux2)=paste(colnames(aux2),"2",sep="_")

aux3=cbind(mat_interact,aux2)
aux3$LOGPRICE=dfgo_imputed$LOGPRICE

dfgo_imputed=aux3


set.seed(123)
index <- sample(nrow(dfgo_imputed),nrow(dfgo_imputed)*0.70)
index_y=which(colnames(dfgo_imputed)=="LOGPRICE" )
go.train <- dfgo_imputed[index,]
go.test <- dfgo_imputed[-index,]

# Criando matrizes de variáveis independentes nas amostras treino e teste
X.train<- as.matrix(go.train[,-index_y])
X.test<-  as.matrix(go.test[,-index_y])

```

```
48
49  #Criando matrizes de variáveis dependentes nas amostras treino e teste
50  Y.train<- go.train[, index_y]
51  Y.test<-  go.test[, index_y]
52
53
54  MQO.fit<- lm(Y.train~X.train)
55  summary(MQO.fit)
56  MQO.fit.coef= as.matrix(MQO.fit$coefficients)
57  X.test_1=cbind(1,X.test)
58  Y_pred.MQO<- X.test_1 %*% MQO.fit.coef
59  RMSE_MQO=sqrt(mean((Y.test-Y_pred.MQO)^2))
60  plot(MQO.fit.coef)
61  coef(MQO.fit)
62
63  cv.ridge<- cv.glmnet(x=X.train, y=Y.train, family = "gaussian", alpha = 0, nfolds = 5)
64  plot(cv.ridge)
65  coef(cv.ridge, s=cv.ridge$lambda.min)
66  lambda_ridge=cv.ridge$lambda.min
67  log_lambda_ridge=log(cv.ridge$lambda.min)
68
69  Y_pred.ridge<- predict(cv.ridge, newx = X.test, s=cv.ridge$lambda.min)
70  RMSE_ridge=sqrt(mean((Y.test-Y_pred.ridge)^2))
71  MSE_ridge=mean((Y.test-Y_pred.ridge)^2)
72
73
74  cv.lasso<- cv.glmnet(x=X.train, y=Y.train, family = "gaussian", alpha = 1, nfolds = 5)
```

```
74  cv.lasso<- cv.glmnet(x=X.train, y=Y.train, family = "gaussian", alpha = 1, nfolds = 5)
75  plot(cv.lasso)
76  coef(cv.lasso, s=cv.lasso$lambda.min)
77  lambda_lasso=cv.lasso$lambda.min
78  log_lambda_lasso=log(cv.lasso$lambda.min)
79
80  Y_pred.lasso<- predict(cv.lasso, newx = X.test, s=cv.lasso$lambda.min)
81  RMSE_lasso=sqrt(mean((Y.test-Y_pred.lasso)^2))
82  MSE_lasso=sqrt(mean((Y.test-Y_pred.lasso)^2))
83
```