PREDIMOV: PREVISÃO DE PREÇOS DE IMÓVEIS

Ramon Vilela[1]; Sidney Carlos Ferrari[2]

Resumo
Os algoritmos de Machine learning vêm se provando um diferencial ao lidar com um grande volume de dados e transformá-los em informação que agregue valor de maneira rápida e eficaz. Este trabalho teve como principal objetivo gerar um algoritmo de modelo de predição do valor de um imóvel de aluguel ou venda, com base em características fornecidas por um usuário através de uma interface, que servirão de insumo para um algoritmo de Regressão Linear Múltipla, ridge e lasso com base nos dados dos imóveis anunciados semanalmente no site do Jornal Negocião, Ourinhos-SP.
Palavras-chave: Machine learning; regressão linear múltipla; modelo de predição.

Abstract
Machine Learning algorithms are proving to be a differential when handling a massive amount of data transforming it into information that aggregates value effectively and quickly. This article has as its main goal to create a price prediction model algorithm for rental or sale property, based on features provided by a user through an interface, that will feed algorithms of Multiple Linear Regression, Ridge, and Lasso, based on weekly real state announced data of on the website of newspaper Negocião, Ourinhos-SP.
Keywords: Machine learning; multiple linear regression; model for predicting.

1 Introdução

O número de dados sendo produzidos por todas as áreas de conhecimento e pela quantidade exorbitante de usuários da internet no mundo aumenta cada vez mais. Devido a isso, surgiu o desafio de transformar todo esse aglomerado de dados em informação. Sobre o campo de ciência de dados, podemos afirmar que:

> A ciência de dados é um campo interdisciplinar que utiliza métodos, processos, algoritmos e sistemas científicos para extrair valor dos dados. Os cientistas de dados combinam uma série de habilidades, incluindo estatísticas, ciência da computação e conhecimento comercial, para analisar dados coletados da web, smartphones, clientes, sensores e outras fontes. A ciência de dados revela tendências e produz as informações que as empresas podem usar para tomar melhores decisões e criar produtos e serviços mais inovadores. Os dados são a base da inovação, mas seu valor vem dos dados de informações que os cientistas podem˜ extrair e depois usar (Oracle, 2022).

O uso dos algoritmos de *Machine Learning* forneceu uma grande ajuda durante a recente pandemia global de Covid-19, auxiliando na tomada de decisões de governos, além de prover insumos que auxiliam no mapeamento e evolução da pandemia, conforme o trabalho

[1] Graduado em Análise e Desenvolvimento de Sistemas pela Faculdade de Tecnologia de Ourinhos-FATEC. E-mail: ramon-vilela@hotmail.com.
[2] Doutor pelo Programa de Pós-Graduação em Engenharia de Produção - PPGEP, campus São Carlos, da Universidade Federal de São Carlos-UFSCAR; professor da Faculdade de Tecnologia de Ourinhos-FATEC. E-mail: sidney.ferrari@fatecourinhos.edu.br.

feito por Carlotto (2021), que utilizou os métodos adaLasso, Lasso, Random Forest e ARIMA para prever o número de infectados em um intervalo de tempo.

Tratando-se do mercado imobiliário, diferente de outros tipos de bens, os imóveis possuem grande variação nos preços, devido aos seus atributos especiais, como imobilidade, durabilidade e seu custo elevado. Como afirmado por González e Formoso (2000), a característica mais importante para estimativa do preço de um imóvel é sua localização, por conta da imobilidade do bem. O valor de localização, por sua vez, está relacionado com a acessibilidade e com as características da vizinhança.

Segundo González e Formoso (2000), existem diferentes procedimentos para realizar a avaliação singular ou coletiva dos imóveis, sendo necessário ponderar os atributos formadores dos valores imobiliários, como a inferência estatística, neste caso, as variáveis julgadas como importantes são reunidas em um modelo de regressão múltipla.

Semanalmente, diversos imóveis para aluguel e venda são anunciados no site Jornal Negocião. Com base nesse grande volume de dados alimentados toda semana, surge a seguinte questão: como prever o valor de imóvel, com base no bairro e como determinar suas características principais para tal previsão? Sendo assim, o objetivo do trabalho foi de desenvolver um algoritmo que realize predições de preços de imóveis através do modelo de regressão linear multiplica, com base nos dados do mercado imobiliário de Ourinhos coletados no site do Jornal Negocião (2022).

2 Metodologia

Os dados utilizados para o algoritmo de previsão, foram retirados do site de anuncio de imóveis da cidade de Ourinhos, o Jornal Negocião, por meio de um algoritmo de coleta desenvolvido na linguagem de programação *Python*, que coletava o nome do bairro, a qual zona pertencia, condição do imóvel (venda ou aluguel), valor e suas características, como quantidade de dormitório, banheiro, WC, garagem, sala, quintal, ´área de serviço, hall, suíte, sala de estar, sala de jantar, sala de tv, edícula, despensa, lavanderia, piscina, copa, tipo (casa ou apartamento), se possuía IPTU, taxas ou valor adicional de condomínio e salvava as informações em um banco de dados MySQL. Os dados foram coletados semanalmente, durante o intervalo de fevereiro de 2019 a março de 2021.

Por conta da disparidade nos valores dos imóveis de venda e aluguel, a massa de dados utilizada para os algoritmos foi dividida com base nessa condição, sendo 11.610 registros para venda e 7.408 registros para aluguel.

Para gerar a predição foram desconsiderados os imóveis do tipo apartamento, por conta da variável "tamanho" no imóvel, que não era mencionada no tipo "casa", tornando assim, inviável uma comparação entre eles, além do número de apartamentos coletados ser muito menor do que o número de casas.

Conforme pontuado por Jardim (2014), o coeficiente de determinação (R^2) é utilizado para estipular em qual percentual a variável dependente (nesse trabalho, o preço do imóvel) é explicada pelas variáveis independentes (demais variáveis), sendo quanto mais próximo de 1, mais preciso o algoritmo se torna. A raiz do erro quadrático médio pega o erro quadrático médio, que apresentaria a informação ao quadrado, e a simplifica, no caso exibindo a variação média do erro do preço do imóvel, tendo como objetivo atingir o valor mais baixo possível (IBM, 2021).

Nos cenários com código por bairro e zona, foi identificado que ao modificar os respectivos códigos, o algoritmo não interpretava como uma variável única, sendo assim, conforme o número que representava o código relacionado aumentava, o valor também se elevava.

Após análise das métricas de zona, ela foi desconsiderada para os demais cenários, devido a disparidade de bairros contidos nessa variável, já que muitos bairros estão dentro de cada zona e cada bairro tem suas particularidades.

Para a Regressão de Lasso, conforme dito por Teixeira (2020), a normalização L1 anula diversas características, e caso não sejam anuladas significa que são importantes. Enquanto a normalização L2 para a Regressão de Ridge escolhe os coeficientes e diminui a variância e diminui o erro, mas não reduz o número de variáveis.

De acordo com Guimarães (2008), as variáveis podem ser divididas em dois tipos: qualitativas e quantitativas. As qualitativas apresentam algum atributo ou qualidade, enquanto as quantitativas apresentam números de uma contagem ou mensuração. Por fim, para o cenário de *One Hot Enc*, foi aplicado o modelo de Regressão Linear Múltipla, entretanto a variável categórica bairro foi transformada em quantitativa por meio da função *One Hot Encoder* da biblioteca *sklearn*, e que cada linha da coluna bairro se tornou uma nova coluna, que deverá receber o valor 1 no bairro que for selecionado pelo usuário, enquanto os demais deverão ter o valor 0 atribuído.

Como forma de avaliar a precisão dos algoritmos, serão considerados os valores do coeficiente de determinação (R^2) e a raiz do erro quadrático médio dos cenários abaixo com suas respectivas variáveis.

• Código Bairro (cod bairro, dormitorio, wc, garagem, quintal, suite, piscina, hall, bairro, lavanderia, despensa, sala, copa, sala de tv, sala de estar, sala de jantar, todos os bairros).
• Código Zona (cod zona, dormitorio, wc, garagem, quintal, suite, piscina, hall, bairro, lavanderia, despensa, sala, copa, sala de tv, sala de estar, sala de jantar, todos os bairros).
• Ridge (cod bairro, dormitorio, wc, garagem, quintal, suite, piscina, hall, bairro, lavanderia, despensa, sala, copa, sala de tv, sala de estar, sala de jantar, todos os bairros).
• Lasso (cod bairro, dormitorio, wc, garagem, quintal, suite, piscina, hall, bairro, lavanderia, despensa, sala, copa, sala de tv, sala de estar, sala de jantar, todos os bairros).
• *One Hot Enc* (dormitorio, wc, garagem, quintal, suite, piscina, hall, bairro, lavanderia, despensa, sala, copa, sala de tv, sala de estar, sala de jantar, todos os bairros).

2.1 Código Fonte

Os dados utilizados nos algoritmos foram coletados pelo código legado abaixo, que percorria as páginas do site do Jornal Negocião, aplicava os filtros com regex em todo o texto da classe que possuia as informações do imóvel.

Figura 1 - Código legado de Coleta dos dados do Jornal Negocião**.**

```
for paginas in range(0,240,20):
    req = Request(link+str(paginas), headers={'User-Agent': 'Mozilla/75.0'})
    fonte = urlopen(req).read()
    sopa  = bs.BeautifulSoup(fonte,'lxml')
    em = sopa.findAll('em')
    for site in sopa.findAll('div', attrs={"class" : "col-md-8 col-sm-8 col-xs-12"}):
        bairro_sem_filtro[contador_imoveis] = site.a.text
        dados = site.text
        bairro_com_filtro[contador_imoveis] = filtro_bairro(bairro_sem_filtro[contador_imoveis])
        imovel[contador_imoveis]= filtro_geral(dados)
        contador_imoveis+=1
```

Fonte: Elaborada pelos autores.

O código abaixo aplica o modelo de Regressão Linear M´Multipla no cenário de código de zona, separa as variáveis de treino e de teste, em seguida gera as informações referentes ao coeficiente de determinação e a raiz do erro quadrático médio.

Figura 2 - Código Regressão Linear Multipla**.**

```
y = dados['preco']
X = dados[['cod_zona','dormitorio','quintal','suite','piscina','hall','
    lavanderia','despensa','sala','copa','sala_de_tv','sala_de_estar','
    sala_de_jantar']]
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size
    =0.3, random_state=3214)

modelo = LinearRegression()
modelo.fit(X_train,y_train)
y_previsto = modelo.predict(X_test)

EQM = metrics.mean_squared_error(y_test,y_previsto).round(2)
REQM = np.sqrt(metrics.mean_squared_error(y_test,y_previsto)).round(2)
R2 = metrics.r2_score(y_test,y_previsto).round(2)
```

Fonte: Elaborada pelos autores.

Abaixo, o código que gera o cenário com *One Hot Enconder*, que transforma cada variável categórica em uma coluna, representada por 0 e 1 de acordo com a linha. Para ambas as condições (aluguel e venda) foram criadas 71 colunas, pois havia 71 bairros diferentes.

Figura 3 - Código *One Hot Encondig*.

```
dados_enc = OneHotEncoder(cols='bairro')
df_negociao_cats = dados[['preco','dormitorio','wc','garagem','quintal'
    ,'suite','piscina','hall','bairro',\
                      'lavanderia','despensa','sala','copa','sala_de_tv'
    ,'sala_de_estar','sala_de_jantar']]
dados_enc = dados_enc.fit_transform(df_negociao_cats)
```

Fonte: Elaborada pelos autores.

3 Resultados

Percebe-se que o cenário que obteve melhores métricas para ambas as condições (aluguel e venda) foi o *One Hot Enc*, enquanto os demais apresentaram pouca variação, sendo assim, as telas de predição de aluguel e venda vão se basear nesse cenário para exibir a predição ao usuário.

Tabela 1 -Métricas de Aluguel.

Cenário	Raiz Erro Quadrático Médio	Coeficiente de determinação (R^2)
Cod Bairro	R$ 370.09	77%
Cod Zona	R$ 359.36	78%
Ridge	R$ 371.49	77%
Lasso	R$ 371.49	77%
One Hot Enc	R$ 274.78	87%

Fonte: Elaborada pelos autores.

Acima, os resultados para os imóveis do tipo aluguel, onde os 4 primeiros cenários apresentaram pequena variação tanto no R^2 quanto na Raiz Erro Quadrático médio, enquanto o cenário *One Hot Enc* apresentou resultados melhores quando comparado aos 4 anteriores, com

destaque para o R^2 que ficou próximo de 87%. Mesmo com as melhorias citadas anteriormente, a Raiz Erro Quadrático Médio persistiu com um valor relativamente alto para o tipo de aluguel.

Tabela 2 - Métricas de Venda.

Cenário	Raiz Erro Quadrático Médio	Coeficiente de determinação (R^2)
Codigo Bairro	R$ 165.347,77	64%
Codigo Zona	R$ 171.701,69	64%
Ridge	R$ 167.265,63	63%
Lasso	R$ 167.015,00	63%
One Hot Enc	R$ 132.027,11	78%

Fonte: Elaborada pelos autores.

Os resultados obtidos nos imóveis do tipo venda apresentaram valores mais imprecisos do que no tipo aluguel, com a Raiz Erro Quadrático Médio alta em todos os cenários, com uma pequena melhoria ao aplicar *One Hot Enc* quando comparado aos demais. Os valores do R^2 se mostraram muito abaixo do esperado. O melhor cenário de 78% foi atingido nos piores aluguéis.

3.1. Interface

Abaixo, a interface para o usuário colocar as informações que serão processadas pelo algoritmo e exibirão o valor aproximado do imóvel, variando de acordo com os parâmetros fornecidos.

Figura 1 - Interface Aluguel.

Fonte: Elaborada pelos autores.

A tela de venda possui a mesma interface de aluguel, com algumas particularidades na escolha do bairro.

Figura 2 - Interface Venda.

Fonte: Elaborada pelos autores.

4 Conclusão

Conclui-se que foi possível realizar a predição do preço dos imóveis, para ambas as condições (aluguel e venda). Entretanto, os valores do coeficiente de determinação e raiz do erro quadrático médio apresentaram valores relativamente elevados, principalmente na condição de venda. Com isso, como trabalho futuro e próximos passos, recomenda-se refazer o código de coleta dos dados do site do Jornal Negocião e aplicar outros algoritmos, como redes neurais, a fim de alcançar valores mais precisos, além de disponibilizar as predições via API para ser integrado a um site.

Referências

CARLOTTO, G. B. **Previsão da evolução da Covid-19 utilizando métodos de Machine Learning**. Trabalho de Graduação (Departamento de Estatística) - Universidade Federal do Rio Grande do Sul-UFRGS, 2021. Disponível em: https://www.lume.ufrgs.br/bitstream/handle/10183/235618/001137217.pdf?sequence=1. Acesso em: 11 set., 2023.

TEIXEIRA, D. M. **Regressão de Lasso, Ridge e Elastic Net.** [S.l.). 1 vídeo (6min:23seg.), 2020. Disponível em: Disponível em:https://www.youtube.com/watch?v=LuttGjfSQCc. Acesso em: 11 set. 2023.

GONZALEZ, M. A. S.; FORMOSO, C. T. **Análise conceitual das dificuldades na determinação de modelos de formação de preços através de análise de regressão**. Minho: Departamento de Engenharia Civil, Universidade do Minho. Disponível em:

https://www.civil.uminho.pt/revista/artigos/Num8/Pag_65-75.pdf. Acesso em: 11 set. 20233.

GUIMARÃES, P. R. B. **Métodos quantitativos estatísticos**. 2 ed. Curitiba [PR]: IESDE Brasil, 2018. Disponível em: https://videoiesde.secure.footprint.net/token=nva=1646678853925~dirs=4~hash=01eb39fe5a37972c1b82c/videoteca/iesde/video/57421_METODOS_QUANTITATIVOS_ESTATISTICOS_2018_PDF/file.pdf. Acesso em: 11 set. 2023.

IBM. **Raiz do erro quadrático médio**. IBM Cloud Pak for Data-IBM, 2021. Disponível em: https://www.ibm.com/docs/pt-br/cloud-paks/cp-data/3.5.0?topic=overview-root-mean-squared-error. Acesso em 11 set., 2023.

JARDIM, S. **Diferença entre o coeficiente de determinação (r^2) e coeficiente de correlação (r)**. [S.l.). 1 vídeo (8min:39seg.), 2014. Disponível em: Disponível em: https://www.youtube.com/watch?v=i5x-UfioHQ4. Acesso em: 11 set. 2023.

JORNAL NEGOCIÃO. **Classificados**. Ourinhos, 2022. Disponível em:https://classificados.negociao.com.br/category/imoveis. Aceso em: 11 set. 2023.

ORACLE. **O que é ciência de dados**. Oracle Cloud Infrastructure, 2022. Disponível em: https://www.oracle.com/br/what-is-data-science/#:~:text=Uma%20plataforma%20de%20ci%C3%AAncia%20de%20dados%20reduz%20a%20redund%C3%A2ncia%20e,e%20incorporando%20as%20melhores%20pr%C3%A1ticas. Aceso em: 11 set. 2023.