Modelo de predição de valores de mercado para apartamentos na Barra da Tijuca utilizando Regressão Linear Múltipla

Marcus Vinicius Ferreira Gonçalves[1], **Carlos Augusto Gomes Xavier**[1]

[1]Escola Nacional de Saúde Pública Sergio Arouca (ENSP)
Fundação Oswaldo Cruz (Fiocruz) – Rio de Janeiro – RJ – Brasil

marcus.goncalves@fiocruz.br, cagxavier@ensp.fiocruz.br

Abstract. *Property appraisal estimates of market value are based on several factors and often occurs in a subjective way not following a well-defined methodology. This paper presents the experimental application of Data Science and Supervised Learning to perform real estate appraisal from data collected on a real state website. The paper describes the steps, results and the construction of a multiple linear regression model for the elaborated context. After completion, the model was tested and presented good predictive ability in relation to the model accuracy.*

Resumo. *A avaliação de imóveis estima o valor de mercado em função de diversos fatores e muitas vezes ocorre de forma subjetiva, complexa e sem uma metodologia definida e de conhecimento público. Este artigo apresenta a aplicação experimental de Ciência de Dados e Aprendizagem Supervisionada para realizar a avaliação de imóveis a partir de dados coletados em um site de compra e venda. O artigo descreve as etapas e resultados e a construção de um modelo de regressão linear múltipla para o contexto elaborado. Após a finalização, o modelo foi testado e apresentou uma boa capacidade preditiva em relação à acurácia do modelo.*

1. Introdução

O mercado imobiliário é responsável por determinar o valor, custos e direitos sobre os imóveis, que através da engenharia de avaliação e seus processos, define tecnicamente o valor do bem. Porém, o valor técnico e o valor de mercado são influenciados pela relação entre o vendedor, o comprador, o imóvel e suas características, o cenário econômico e a especulação imobiliária.

O valor de mercado de um imóvel é normatizado pela NBR 14653-1 [ABNT 2001], sendo os imóveis urbanos normatizados pela NBR 14653-2. As normas definem a classificação dos bens, as atividades, os aspectos quantitativos e qualitativos, a situação mercadológica, a identificação do valor de mercado em função da metodologia e métodos de estimar, entre outros. Definem, também, os itens necessários para a apresentação do laudo de avaliação e que este laudo deve ser realizado por um engenheiro de avaliações. Todavia, não deve ser confundido com o preço, que corresponde à quantia paga pelo comprador ao vendedor.

A NBR 14653-1 [ABNT 2001] preconiza que deve ser observado o maior número de imóveis com o máximo de atributos semelhantes como forma de estimar o valor de mercado, utilizando comparação entre imóveis e buscando a maior similaridade entre

eles. Utilizando esta normativa, é apropriado o uso de técnicas de ciência de dados, o que é proposto para este estudo de caso.

Na literatura científica existem trabalhos e revistas que tratam o assunto da avaliação imobiliária. Muitos destes trabalhos apresentam técnicas utilizando aprendizagem de máquina, realizadas a partir de dados provenientes de bancos de dados ou documentações imobiliárias, que passam pelas etapas de ciência de dados. Como trabalhos relacionados [Nunes et al. 2019][Neto 2006] utilizam Regressão Linear Múltipla como modelos para permitirem a avaliação imobiária.

Este estudo de caso descreve a aplicação e experimentação de técnicas na construção de um modelo de predição de valor de mercado de imóveis através de comparação, utilizando o algoritmo de Aprendizagem de Máquina de Regressão Linear Múltipla, a partir da entrada de dados correspondentes aos valores das características dos apartamentos novos e usados, excluindo lançamentos, anunciados no site ZAP Imóveis[1] em 13 de agosto de 2019 e localizados na Barra da Tijuca na cidade do Rio de Janeiro.

Para o estudo de caso são utilizadas as ferramentas de tecnologia, ciência de dados e aprendizado de máquina: Microsoft Excel, Python[2], Numpy[3], Pandas[4], Matplotlib[5] e Scikit-learn[6]. A seguir são apresentadas as etapas do estudo.

2. Estudo de caso

Os dados nos anúncios dos imóveis foram listados no site ZAP Imóveis, através de uma busca simples de compra de apartamento padrão no bairro da Barra da Tijuca na cidade do Rio de Janeiro. No dia 13 de agosto de 2019 foi realizada a consulta e coleta. O site apresentou um total 17.632 apartamentos para venda, distribuídos em 916 páginas, cada uma contendo 3 lançamentos sem valor e 22 apartamentos com dados disponíveis. Os anúncios continham fotos e as informações dos imóveis: valor de venda, valor do IPTU, valor do condomínio, número de quartos, banheiros, vagas, área útil em metros quadrados, localização, cidade, estado e descrição do imóvel.

Para compor a amostra foram escolhidas as primeiras 22 páginas, que traziam os anúncios mais vistos e buscados ou patrocinados, totalizando 1078 apartamentos, pouco mais de 6% do total listado. Além do percentual, o motivo da escolha deste número de corte se deve ao fato das páginas seguintes apresentarem muitos imóveis repetidos e já listados, assim como imóveis com menos informações, inclusive sem as mais relevantes.

2.1. A coleta e limpeza dos dados

Para a coleta foi desenvolvido um pequeno trecho de código em Python, que retornou com as informações dos 1078 apartamentos apresentados na busca. O código separou estas informações por campos e armazenou em um arquivo texto de extensão .CSV, que significa valores separados por vírgula (*Comma-Separated Values*). A limpeza removeu anúncios do mesmo imóvel em diferentes corretoras, anúncios mal preenchidos: valores

[1]Site do ZAP Imóveis: site de compra e venda de imóveis. http://zapimoveis.com.br
[2]Python v.3.7.4: Linguagem de programação orientada a objeto e de alto-nível. http://python.org
[3]Numpy v.1.17.0: Pacote Python com funções matemáticas para matrizes e arranjos. http://numpy.org
[4]Pandas v.0.25.0: Biblioteca Python para manipulação e análise de dados. http://pandas.pydata.org
[5]Matplotlib v.3.1.1: Biblioteca Python de plotagem de gráficos. http://matplotlib.org
[6]Scikit-learn v.0.21.3: Biblioteca Python de Aprendizagem de Máquina. http://scikit-learn.org

errados ou faltantes (*missing*) e anúncios que não pertenciam ao bairro. Ao final da etapa, restaram 880 apartamentos, aproximadamente 5% do total disponível para coleta.

2.2. Análise dos dados

A primeira análise do conjunto de dados resultante foi verificar a dispersão dos valores dos apartamentos e descarte de algumas variáveis não funcionais. Para analisar as variáveis independentes foi esboçado um diagrama de caixa (box-plot) para cada variável, permitindo identificar os valores discrepantes (*outliers*). O passo seguinte foi analisar a relação entre a variável dependente (valor do imóvel) e cada uma das variáveis independentes, assim como realizar o estudo da Correlação de Pearson entre todas as variáveis, para verificar a existência de multicolinearidade, o que seria prejudicial na execução do algoritmo escolhido de Regressão Linear Múltipla [McKinney 2019]. De forma a resultar em quatro variáveis independentes: número de quartos, de banheiros, de vagas e metros quadrados. Todas elas relacionadas com o valor do imóvel.

2.3. Construção, teste e análise do modelo

Após a etapa de análise foi iniciada a etapa de construção do modelo utilizando Regressão Linear Múltipla. O objetivo da Regressão Linear Múltipla é desenvolver a equação modelo, a partir do treinamento das variáveis independentes [McKinney 2019]: número de quartos, número de banheiros, número de vagas e número de metros quadrados, em relação à variável dependente valor do imóvel.

Para a construção do modelo de regressão com abordagem supervisionada, a amostra foi dividida em dois conjuntos: treinamento e teste. Foi utilizada uma função do Scikit-learn que gerou aleatoriamente os dois conjuntos. O conjunto de treinamento ficou com 67% do número total da amostra, o equivalente a 589 apartamentos e o conjunto de treino ficou com 33%, equivalente a 291 apartamentos.

O modelo de Regressão Linear Múltipla foi construído (treinado) a partir do conjunto de treinamento e logo em seguida o conjunto de testes foi submetido ao modelo. Para a análise do modelo de regressão foram esboçados: o gráfico de resíduos para os valores preditos para o conjunto de treinamento e teste e o histograma da distribuição de variação de resíduo de treinamento e teste.

A partir dos gráficos de resíduos de treinamento e teste foi possível verificar que os valores dos resíduos estavam distribuídos aleatoriamente ao redor de uma linha que parte da origem, indicando uma variância constante. Neste gráfico também foram observados os resíduos discrepantes gerados na predição. Já no gráfico do histograma da variação do resíduo foi verificado que a distribuição é concentrada e normal, localizando a maior frequência próximo ao valor zero. Os gráficos da distribuição normal dos resíduos confirmaram esta afirmação para os dois conjuntos de dados: treinamento e teste.

O sumário estatístico foi gerado e apresentou o resultado do coeficiente de determinação R2 ou coeficiente de explicação do modelo. Este coeficiente é uma medida do modelo linear generalizado em relação aos valores observados. É utilizado para regressões lineares e seu valor varia entre 0 e 1, que pode ser associado a valores percentuais [Mann 2007]. O valor de R2 para o conjunto de treinamento foi de 0,77 e para o conjunto de teste de 0,79, o que significa uma taxa de explicação do modelo de 77% para treinamento e 79% para teste.

2.4. Predição utilizando o modelo e comparações

Com o modelo construído foram realizados sete novos testes com valores determinados. Foi criada uma tabela com 7 imóveis fictícios, cujos dados estivessem distribuídos dentro dos limites das variáveis do modelo, possibilitando montar sete segmentos de respostas esperadas. Os valores das variáveis e resultados de predição para estes testes gerados pelo modelo apresentaram valores coerentes variando resultados entre -13% e 14% da meta. Apenas um teste apresentou erro de 44% acima, porém este teste encontrava-se no conjunto de dados (*outliers*), que apresentava os maiores valores discrepantes.

O site do Imóvelweb[7] oferece um serviço chamado de Precificador. Este serviço gera para o usuário uma faixa de valores de imóveis a partir das características escolhidas. Ele funciona através da comparação de imóveis que estão na mesma localização, possuem características similares e estão publicados na base de dados do site. Foram submetidos os mesmos valores de entrada dos sete novos testes e os resultados se encaixaram dentro do limite do Precificador. Apenas um dos testes apresentou 21% de erro.

A partir dos testes executados, o modelo de predição utilizando Regressão Linear Múltipla foi satisfatório, demonstrando ser genérico o sufiente para o conjunto de dados e o número de variáveis selecionadas, evitando sobre e sub-ajustes (*overfiting* e *underfitting*) e apresentando percentual de explicação do modelo de 79% com a base coletada.

3. Conclusão de discussão dos resultados

Este trabalho evidenciou métodos e possibilidades da utilização da Regressão Linear Múltipla para estimar valores de imóveis, desde que sejam realizados todos os procedimentos com os dados, principalmente o estudo estatístico das variáveis para evitar a multicolinearidade. Após o modelo criado foram realizados testes que permitiram comparar os dados e validar o modelo. Mesmo com uma taxa de explicação do modelo - R2 de 79%, foi considerado um desempenho aceitável para a amostra coletada de 5% do que estava disponível e para apenas quatro variáveis preditoras (independentes).

A utilização do modelo construído restringe-se apenas à apartamentos do bairro da Barra da Tijuca em função da particularidade dos dados que originaram este, porém nada impede a aplicação da metodologia utilizada para a construção de um outro modelo para outra localidade, além de aplicação de diferentes modelos de regressão.

Referências

ABNT (2001). *NBR 14653-1: avaliação de bens- parte 1: procedimentos gerais*.

Mann, P. S. (2007). *Introductory statistics*. John Wiley & Sons.

McKinney, W. (2019). *Python para análise de dados: Tratamento de dados com Pandas, NumPy e IPython*. Novatec Editora.

Neto, A. P. (2006). Redes neurais artificiais aplicadas às avaliações em massa estudo de caso para a cidade de belo horizonte/mg.

Nunes, D. B., Barros, J. d. P., and Freitas, S. M. d. (2019). Modelo de regressão linear múltipla para avaliação do valor de mercado de apartamentos residenciais em fortaleza, ce. *Ambiente Construído*, 19:89–104.

[7]Site Imóvelweb: site de anúncio de imóveis e do Precificador. http://imovelweb.com.br