Escrever título (escolher no final)

Universidade Federal da Paraíba - CCEN

Gabriel de Jesus Pereira

14 de junho de 2024

Índice

1 Resumo

2 Capítulo 1

– Fazer antes da conclusão –

2.1 Introdução

2.2 Objetivos

2.2.1 Objetivo Geral

2.2.2 Objetivos Específicos

2.3 Organização do Trabalho

3 Capítulo 2

– Fazer depois da metodologia –

3.1 Recursos Computacionais

3.1.1 Linguagem de Programação R

3.1.2 Quarto

3.1.3 Linguagem de Programação Python

3.1.4 Web Scraping

4 Metodologia

4.1 Os dados e o procedimento adotado para sua obtenção

O Web scraping, também conhecido como extração de dados da web, é uma técnica utilizada para o processo de coleta de dados estruturados da web de maneira automatizada. É um processo que vem sendo constantemente utilizado por instituições públicas e privadas para a construção de produtos que utilizam algoritmos de aprendizagem de máquinas, observa ofertas e discontos, faz análise de mercado ou monitoração de marcas.

Neste projeto, para fins de estudo e análise do mercado imobiliário, os dados foram coletados por meio de extração de dados do site do Zap Imóveis. O Zap Imóveis é um site do Grupo OLX que reúne ofertas do mercado imobiliário e que funciona como uma plataforma dinâmica para facilitar a conexão entre quem deseja alugar, comprar ou vender um imóvel; podendo servir também para corretores ou outros profissionais do setor de imóveis. Este projeto foi possível graças as informações que foram coletadas do site do Zap Imóveis em dois diferentes períodos do ano de 2023. O primeiro deles, as informações foram coletadas utilizando variados pacotes para raspagem de dados e proxies rotativas da linguagem de programação R, a fim de evitar ser bloqueado pelos mecanismos de segurança do site. Na segunda etapa, os dados foram coletados empregando a linguagem de programação Python com as bibliotecas Scrapy e Playwrite, que serve para web crawling e web scraping, e o Playwrite que serve para testes em aplicativos da web, mas que neste caso foi utilizado para manejar páginas dinâmicas.

Desta forma, com a ideia de modelar o valor do imóvel e analisar o mercado imobiliário, foram coletados aqueles variáveis que estavam disponíveis no site do Zap Imóveis e que poderiam de alguma forma ser significativas ao tentar explicar o valor do imóvel durante a sua modelagem. Assim, no total foram coletatas 23 variáveis, das quais 8 são quantitativas e 15 qualitativas nominais, sendo 13 de caráter dicotômico. No entanto, nem todas essas variáveis foram coletadas diretamente do Zap Imóveis, a latitude e longitude foram obtidas pela geocodificação do endereço utilizando o pacote tidygeocoder da linguagem de programação R. Portanto, temos as seguites variáveis:

- Valor do imóvel: esta é a variável dependente, aquela que será modelada e será o principal objeto de estudo deste trabalho;
- Área: área do imóvel em m^2;
- Condomínio: valor pago pelo condomínio;
- IPTU: imposto cobrado de quem tem um imóvel urbano;
- Banheiro: quantidade de banheiros presentes na propriedade;
- Vaga de estacionamento: quantidade total de vagas de estacionamento;
- Quarto: quantidade de quartos no imóvel;
- Latitude: posição horizontal medida em frações decimais de graus;
- Longitude: posição vertical que, assim como a latitude, é medida em frações decimais de graus;
- Tipo do imóvel: foram obtidos 7 tipos de imóveis, apartamentos, casas, casas comerciais, casas de condomínio, casas de vila, coberturas, lotes comerciais e de condomínio;
- Endereço: nome do endereço do imóvel;
- Variáveis dicotômicas que indicam se o imóvel tem ou não aquela característica (representado como 1 ou 0, respectivamente): área de serviço, academia, elevador, espaço gourmet, piscina, playground, portaria 24 horas, quadra de esporte, salão de festa, sauna, spa e varanda gourmet.

No entanto, devido a observações feitas durante o estudo, nem todas essas variáveis foram utilizadas para a modelagem do valor dos imóveis, seja por conter muitos valores valores ausentes ou por não ter se mostrado significante para o que se desejava explicar. Ainda, como a coleta destes dados foram feitas em dois momentos distintos, temos dois bancos de dados, um com 29712 observações e o outro com 14956. Por fim, essas duas bases de dados foram unidas e, para não correr o risco de conter imóveis repetidos, aqueles que tinham o mesmo número de identificação foram removidos.

4.2 Descritiva dos dados

A análise exploratória de dados marca uma das primeiras etapas de qualquer estudo que utiliza a estatística como uma de suas principais ferramentas, pois permite encontrar padrões de comportamento no dados, descobrir relações entre as variáveis estudadas. Dessa forma, a primeira etapa desse estudo, após a coleta e organização dos dados obtidos do Zap Imóveis, foi fazer uma descritiva dos dados. Essa etapa permitiu encontrar padrões nos diferentes tipos de imóveis bem como o seu tipo pode influenciar na características do imóvel, o que, por consequência, pode afetar o seu valor. Assim, para identificar esses diferentes comportamentos, foram criados gráficos e tabelas a fim de caracterizar as relações das variáveis independentes com a dependente.

4.3 Aprendizado Supervisionado e não supervisionado

Na aprendizagem de máquinas, uma das estapas mais importantes é saber qual técnica será utilizada para resolver um problema que se enquadra em diferentes formas de aprendizado. Para isso, existem mais de uma forma em que um algoritmo consegue utilizar os dados e explicar o que está sendo modelado a partir deles. No entanto, a maioria dos problemas de aprendizado de máquinas recais em dois casos mais conhecidos: aprendizado supervisionado e não supervisionado.

4.3.1 Aprendizado supervisionado

Suponha uma regressão logística. Sabemos que na regressão logistíca temos um modelo com a seguinte forma $Y_i = f(X) + \epsilon$, em que Y_i assume 0 ou 1 para classificar o que está sendo modelado e representa a variável dependente, $f(X)$ representa as variáveis independentes que serão utilizadas para a modelagem e ϵ representa o erro da regressão. Dessa forma, podemos considerar o caso em que a regressão logística tenta classificar pacientes que podem ou não estar com diabetes. Para isso, utilizariamos variáveis significativas para a classificação do estado de cada paciente. Esse exemplo é conhecido como aprendizagem supervisionada. Na aprendizagem supervisionada, busca-se aprender Y_i atráves de um exemplo. Nesse caso, as variáveis dependentes podem ser interpretadas como o exemplo, as informações de relações de pacientes que podem ter ou não diabetes, e o estado do paciente pode ser interpretado como o que se deseja aprender. Este processo é entendido como *aprendizado por exemplo*, HASTIE *et al.* (2009). O aprendizado supervisionado pode aparecer em casos de regressão linear, regressão logística, ou até mesmo em métodos mais modernos, como GAM, boosting e máquina de vetores de suporte, JAMES *et al.* (2023).

4.3.2 Aprendizado não supervisionado

Por outro lado, o aprendizado não supervisionado aparece em situações mais desafiadores, pois não há um exemplo para explicar aquilo que se pretende explicar. Este processo é conhecido como *aprendizado sem exemplo*, HASTIE *et al.* (2009). Dessa forma, no aprendizado não supervisionado, tem-se uma amostra com N observações $(x_1, ..., x_N)$ de um vetor aleatório X com densidade conjunta $f(x)$ em que o objetivo é inferir propriedades da densidade sem ajuda de exemplos para cada observação. Assim, como há uma falta de uma variável resposta y_i para supervisionar a análise, pode-se procurar entender a relação entre as variáveis ou as observações, JAMES *et al.* (2023). Por exemplo, uma das técnicas mais aplicadas em problemas que envolvem o aprendizado supervisionado é a análise de cluster, em que o objetivo é determinar, com base em $x_1, ..., x_n$, se as observações são caracterizadas em grupos distintos. Esse é um dos métodos que poderiam ser aplicados, por exemplo, na análise de crédito de clientes de um cartão de crédito, tornando possível analisar o seu perfil e classificá-lo em diferentes grupos para recomendar produtos especificos adequados ao seu perfil.

4.4 Métodos de reamostragem

4.5 Tunagem de hiperparâmetros com otimização bayesiana

Na aprendizagem de máquina, uma das principais etapas é a tunagem do hiperparâmetros, que se consiste em encontrar a melhor combinação de hiperparâmetros de um modelo. A tunagem desses hiperparâmetros torna a maioria dos algoritmos de aprendizagem de máquina configuráveis, permitindo encontrar modeloos com resultados que conseguem melhores generalizações. No entanto, essa configuração vem de forma não trivial.

Existem diversas técnicas para otimização de hiperparâmetros utilizadas em aprendizagem de máquina. Uma das técnicas mais comuns é a GridSearch, que se propõe em testar todas as combinações possíveis previamente definidas num espaço de procura e escolher aquela que obteve os melhores resultados. No entanto, a GridSearch

e muitas outras técnicas tem performances abaixo do esperado devido ao seu alto custo computacional, possibilitando testar apenas uma pequena de hiperparâmetros e tendo, por efeito, modelos com resultados piores quando comparados com aqueles que foram otimizados com métodos mais eficientes e com melhores estratégias de procura.

4.5.1 Otimização Bayesiana

4.5.2 Tree-Structured Parzen Estimator

O Tree-Structured Parzen Estimator (TPE) é uma variante de métodos otimização bayesiana. A otimização bayesiana tem como finalidade minimizar uma função objetivo definida da seguinte forma:

$$x_{opt} \in \underset{x \in \chi}{\operatorname{argmin}} f(x)$$

Como foi dito anteriormente, a otimização de hiperparâmetros na aprendizagem de máquinas busca encontrar a combinação de hiperparâmetro x_{opt} que produza os melhores resultados, como, por exemplo, a configuração com o menor erro quadrático médio em problemas de regressão. A procura por x_{opt} é feita a partir de uma função da família de funções de aquisição.

Na prática, a otimização bayesiana tem como objetivo indireto otimizar uma *função de aquisição*. As funções de aquisição tendem a ser computacionalmente mais eficientes para estimação com gradientes analiticamente manejáveis, permitindo o uso de otimizadores prontos para estimar cada observação GARNETT (2023).

4.6 Modelos baseados em árvores

5 Capítulo 4

5.1 Resultados

6 Capítulo 5

6.1 Conclusão e Discussão

7 Referências

GARNETT, R. **Bayesian optimization**. [S.l.]: Cambridge University Press, 2023.

HASTIE, T. *et al.* **The elements of statistical learning: data mining, inference, and prediction**. [S.l.]: Springer, 2009. V. 2.

JAMES, G. *et al.* **An introduction to statistical learning: With applications in python**. [S.l.]: Springer Nature, 2023.