

Um modelo híbrido de clusterização e classificação para previsão de preços de aluguéis de imóveis em cidades litorâneas

Daniel Shinoda Pascoal¹

Universidade de São Paulo

Victor Claudio Bento Camargo²

Universidade Federal de São Carlos

1 Introdução

Prever preços de imóveis utilizando algoritmos de *machine learning* a partir de dezenas de variáveis explicativas é uma tarefa conhecida dentro da comunidade de ciência de dados. Nesse estudo, o objetivo é focar na realidade de imóveis de cidades litorâneas no Brasil, mais especificamente os apartamentos da Região da Zona Sul do Rio de Janeiro que é conhecida por grandes contrastes na renda da população. Fatores como proximidade com comunidades carentes e distância à praia podem fazer o preço do aluguel ter grande variação dentro de um mesmo bairro. Dentro desse contexto, pretendemos apresentar os passos que vão desde a coleta até a modelagem de um modelo híbrido de clusterização e classificação, com o objetivo de conseguir a melhor estimativa de preços de aluguel a partir de dados publicamente disponíveis. Esse modelo de aprendizado de máquina, uma vez treinado pode ser utilizado para fornecer estimativas fidedignas para interessados em obter o preço praticado na região para imóveis similares ao seu. Técnicas de *web scrapping* em um *site* de anúncios de imóveis foram usadas para fazer a coleta de dados, do qual podemos obter dados de 3942 imóveis e 8 variáveis explicativas. Além disso, através de *feature engineering* na variável endereço, podemos conseguir a distância do imóvel à praia. Para a regressão, foram comparados o modelo de *random forest* e um modelo híbrido de clusterização e *random forest*.

2 Resultados e Discussão

Foram coletados dados de um site de anúncio de imóveis da região do Rio de Janeiro, dos bairros de Copacabana, Ipanema e Leblon. A coleta foi completa de todos os imóveis utilizando

¹daniel.shinoda@usp.br

²victor.camargo@dep.ufscar.br

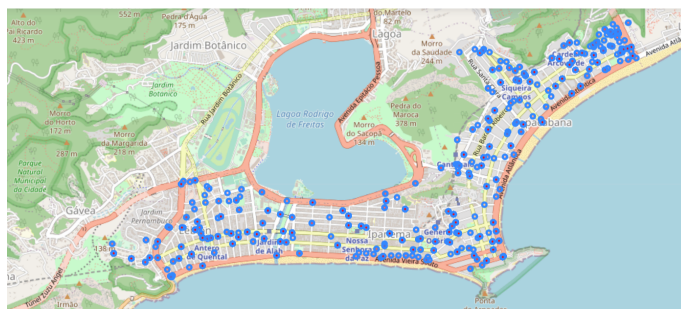


Figura 1: Mapa da zona sul do Rio de Janeiro e dados coletados.

o mecanismo de busca do *site*. Nosso conjunto de dados é composto por 3942 imóveis e os dados obtidos por *web scrapping* compreendem valores para as variáveis: preço do aluguel, valor do condomínio, área (m^2), número de vagas na garagem, número de banheiros, número de quartos, endereço e bairro. Um mapa da localização dos imóveis na região da zona sul pode ser visto na Figura 1. A amostra cobre uma porção significativa dos bairros, indo desde a proximidade da praia até os pontos mais afastados.

2.1 Coleta de Dados e Pré-Processamento

Na coleta de dados, encontramos 1648 valores nulos na variável vagas, 141 na variável valor de condomínio e 4 na variável banheiro. Como muito dos prédios da região não possuem vagas de estacionamento, adicionamos o valor zero. No caso dos banheiros, assumimos impossível o aluguel de um apartamento sem banheiros e adicionamos o valor 1. Já no caso dos condomínios, nossa abordagem foi a adição da mediana dos valores de condomínio do bairro em questão. Alguns imóveis foram removidos da amostra por terem preços de aluguéis ou valores de condomínio muito elevados, ao qual julgamos serem erros de digitação.

2.2 Análise Exploratória

Os histogramas apresentados na Figura 2 mostram que os dados da região possuem distribuição assimétrica positiva de preço do aluguel. Foram analisados os gráficos de dispersão entre todas as variáveis numéricas, onde foi observada uma leve tendência de correlação positiva entre preço do aluguel e área e, entre preço do aluguel e valor de condomínio.

2.3 Feature engineering

Intuitivamente, a distância do imóvel à praia pode ser uma variável explicativa. Para obter essa informação, utilizamos os quiosques de praia. Esses estabelecimentos comerciais (peculiares das praias do Rio de Janeiro) contornam quase que todo o litoral dessa região, e em alguns pontos também utilizamos a localização de imóveis residenciais localizados na praia. O endereço desses pontos foram utilizados como *proxies* para obter a distância do imóvel à praia. A latitude e longitude dos pontos foram obtidos via API do Google Maps [4], de posse da latitude e longitude

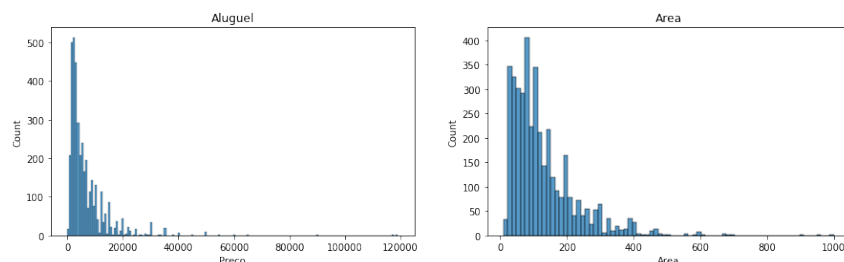


Figura 2: Histogramas de preço de aluguel e área.

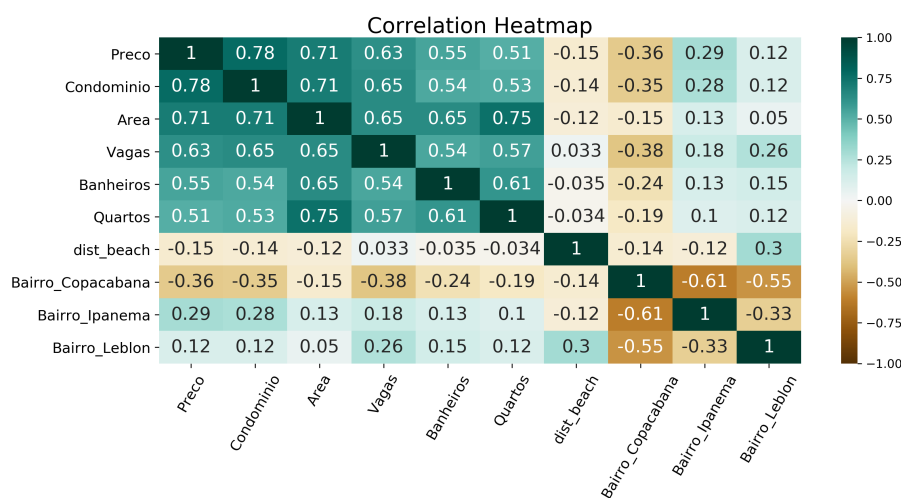


Figura 3: Heatmap de correlação das variáveis numéricas.

foram calculadas as distâncias entre todos os imóveis aos *proxies*, e dentre essas diferenças a menor distância é a escolhida como a distância da praia daquele imóvel. Para cálculo da distância foi utilizado a fórmula Vincenty [3] que assume a figura da terra como um esferoide achatado.

2.4 Modelagem

Foram criadas variáveis *dummy* para as variáveis categóricas de bairro obtendo mais 3 features com valores esparsos. A Figura 3 acima mostra o mapa de calor de correlação entre todas as variáveis. Podemos, principalmente, notar correlação positiva da variável preço do aluguel com as seguintes variáveis explicativas: valor do condomínio (0,78), área (0,71) e vagas (0,63). Aplicamos o algoritmo de regressão *random forest* [2] com o método de otimização de parâmetros conhecido como *random search* [1], em que, diversas variáveis explicativas são testadas para encontrar aquelas que mais influenciam no resultado de previsão. Com as melhores variáveis, obtemos um erro médio percentual absoluto (MAPE) total de 32,9% e um erro médio absoluto de R\$ 1.768,00. Vale ressaltar o erro médio absoluto também é influenciado pela presença de imóveis com valor elevado, aplicando o modelo apenas em apartamento de preço inferior a R\$ 5.000,00, nosso erro médio absoluto foi de R\$ 787,00. Uma vez que existem diversos tipos de imóveis com carac-

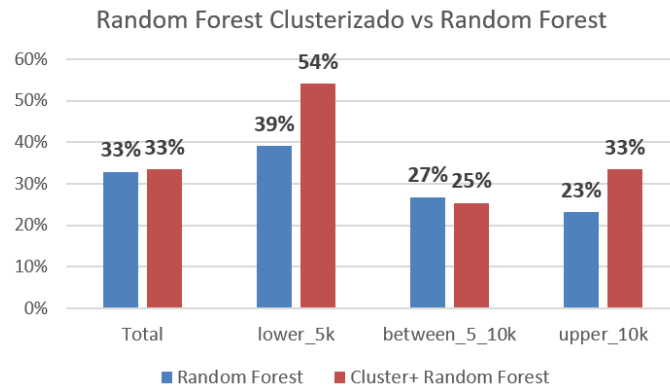


Figura 4: MAPE nas 3 faixas de preço.

terísticas e preços tão distantes nessa região, optamos por aplicar uma clusterização utilizando o algoritmo kmeans para depois aplicar nosso algoritmo otimizado de *random forest* em cada um dos 3 *clusters*. O MAPE total foi de 33,4%, obtendo um pior resultado para imóveis nos perços de alugueis abaixo de R\$ 5.000,00 e acima de R\$10.000,00 e levemente melhor nas faixas entre R\$ 5.000,00 e R\$10.000,00. A Figura 4 resume o resultado em 3 faixas de preço, até R\$ 5.000,00, de R\$ 5.000,00 até R\$ 10.000,00 e de R\$ 10.000,00 em diante.

3 Conclusões

O modelo híbrido de clusterização e classificação para previsão de preços de alugueis de imóveis em cidades litorâneas apresentado nesse trabalho obtém um erro médio percentual absoluto de 32,9% e de 39% nas faixas até R\$5.000,00, 27% nas faixas de R\$5.000,00 a R\$10.000,00 e 23% nas faixas acima de R\$10.000,00. Aplicar o modelo de kmeans nas variáveis não trouxe resultados melhores nas faixas, com exceção da faixa de R\$5.000,00 a R\$10.000,00. Em estudos futuros pretendemos aplicar o algoritmo de clusterização junto a técnicas de redução de dimensionalidade. Além disso, outros algoritmos como *support vector machines* e redes neurais podem trazer resultados melhores de previsão. Pretendemos ainda encontrar uma forma de medir a distância da residência a comunidades de risco do Rio de Janeiro.

Referências

- [1] J. Berstra, Y. Bengio. Random Search for Hyper-Parameter Optimization, *Journal of Machine Learning Research*, 13:281–305, 2012.
- [2] L. Breiman. Random forests, *Machine Learning*, 45:5–32, 2001.
- [3] T. Vincenty. Direct and Inverse Solutions of Geodesics on the Ellipsoid with application of nested equations, *Survey Review*, 1975. DOI: 10.1179/sre.1975.23.176.88
- [4] <https://developers.google.com/maps/documentation>