Grupo ZAP: Data Science Challenge

Autor: Rodrigo de Lima Oliveira
Linkedin: https://www.linkedin.com/in/rodrigolima82/

Visão Geral

Desafio: Quanto custa?

- Dados: anúncios fornecidos pelo Grupo ZAP
- Objetivo: Automatizar o processo de estimativa do **preço de venda de apartamentos**
- Resumo dos Dados
- Construção de Variáveis
- Seleção de Atributos
- Metodologia e Resultados do Modelo
- Considerações Finais
- Respostas ao desafio

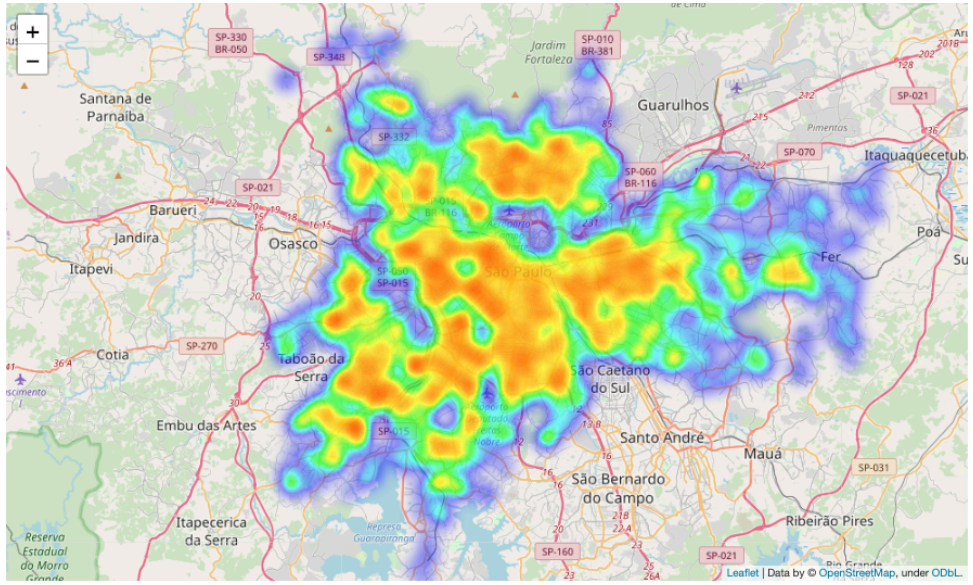
Resumo dos Dados

Total de registros: 133.964

28 atributos

Total de apartamentos para venda no dataset de treino: 64.146

Detalhe dos Dados



- O gráfico apresenta a distribuição dos apartamentos à venda na cidade de São Paulo

- O mapa de calor indica o preço de venda dos apartamentos (quanto mais vermelho, mais alto o valor)

Construção de Variáveis

HASPARKINGSPACES: COLUNA CRIADA PARA INDICAR SE O IMÓVEL TEM ESPAÇO PARA ESTACIONAMENTO

HASBATHROOMS: INDICA SE O IMÓVEL TEM PELO MENOS 1 (UM) BANHEIRO

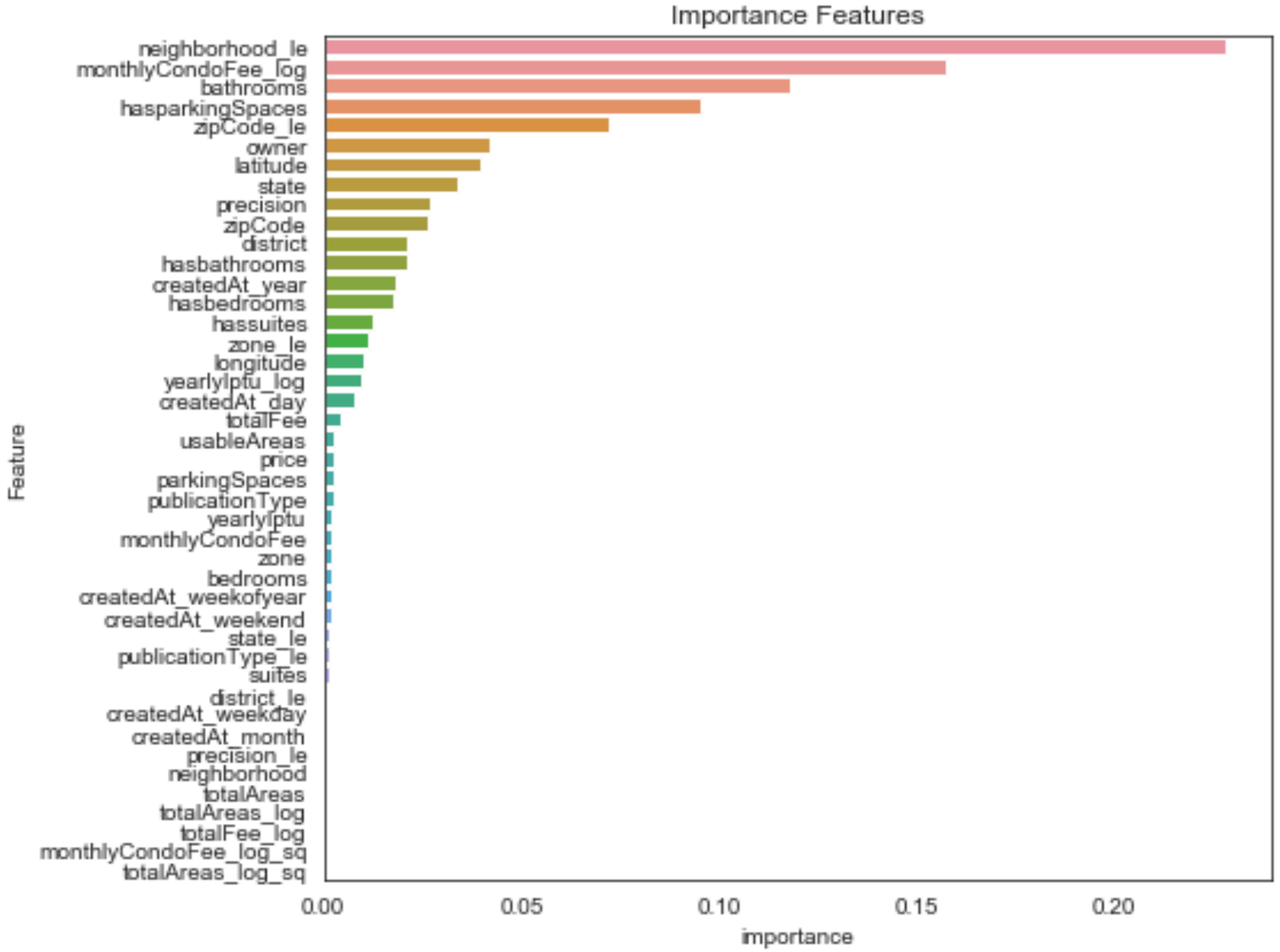
HASBEDROOMS: INDICA SE O IMÓVEL TEM PELO MENOS 1 (UM) QUARTO

HASSUITES: INDICA SE O IMÓVEL TEM PELO MENOS 1 (UMA) SUITE NO APARTAMENTO

TOTALFEE: SOMATÓRIA DO VALOR DA TAXA DO CONDOMÍNIO E O VALOR DO IPTU

Seleção de Atributos

- Colunas selecionadas com base na importância
- (aplicado técnica de Feature Selection)



Métrica

Métrica para avaliação do modelo

Root Mean Squared Error (RMSE)

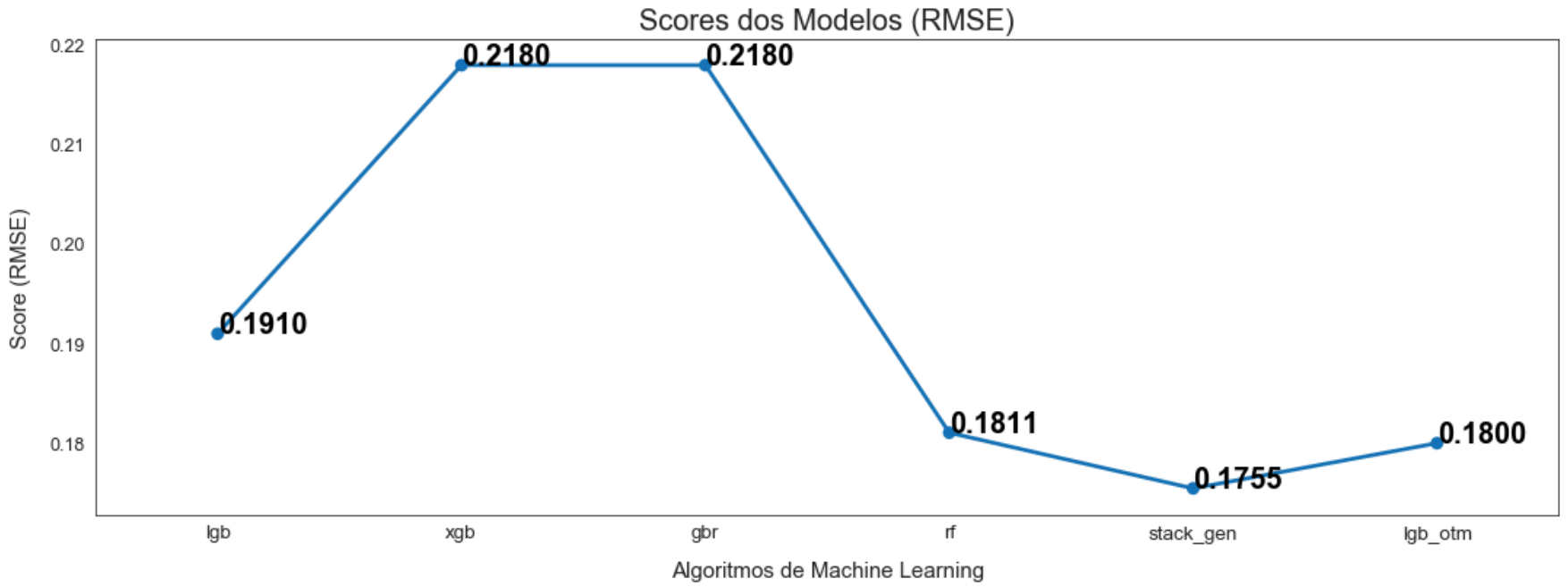
- Essa é uma excelente métrica para modelos de regressão, além de ser muito fácil de interpretar.
- A **Raiz Quadrada do Erro Quadrático Médio —** nada mais é que a diferença entre o valor que foi previsto pelo modelo e o valor real que foi observado
- No nosso projeto onde o modelo estima o preço dos apartamentos.
- O modelo deveria ter estimado o valor de R$ 100 mil (exemplo), mas ele estimou R$ 99 mil: **esse -R$ 1 mil de diferença é o erro do modelo**.
- Então repete-se esse processo para todo o conjunto de dados, **eleva-se o erro ao quadrado**, **tira-se a média** de todos os valores do conjunto e, por fim, **calcular a raiz quadrada.**

Modelos

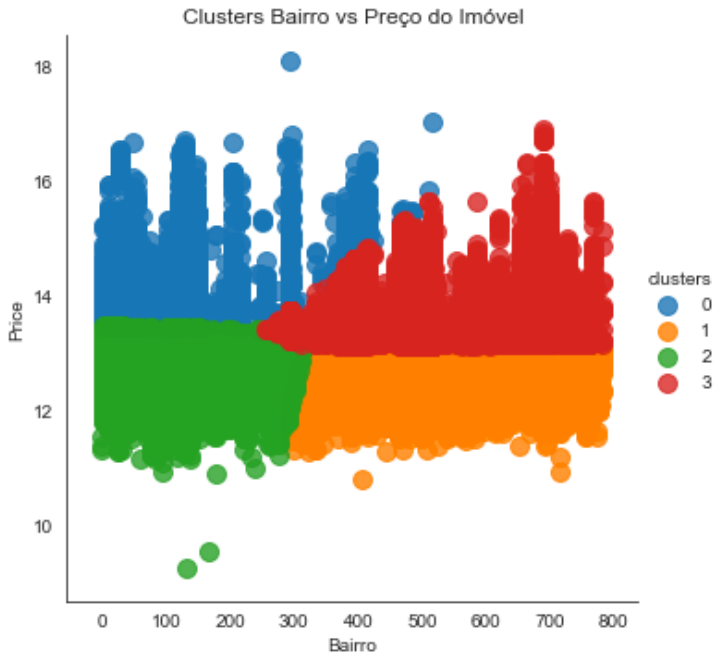
Modelos de Machine Learning avaliados:

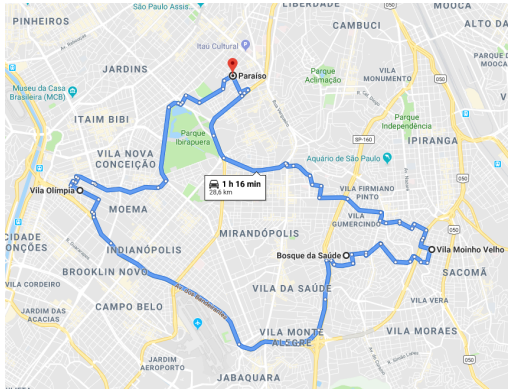
- Light Gradient Boosting (lgb)
- XGBoost (xgb)
- Gradient Boosting (gbr)
- Random Forest (rf)
- Combinando os modelos (stack_gen)
- Light Gradient Boosting (otimizado)

Resultado dos modelos



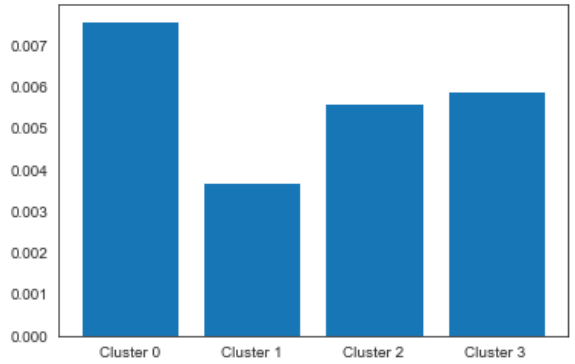
Métrica RMSE	Treino	Teste
StackingCVRegressor (combinação de algoritmos) * melhor performance**	0,1755	0,1789





Bairros centróides dos clusters

Cluster 0: Paraiso
Cluster 1: Vila Moinho Velho
Cluster 2: Bosque da Saude
Cluster 3: Vila Olimpia



Métrica RMSE

Cluster 0: 0.0076
Cluster 1: 0.0037
Cluster 2: 0.0056
Cluster 3: 0.0059

Resultado por cluster
Bairro x Preço

Considerações Finais

Possibilidade de melhorar a performance aumentando a quantidade de k-folds (avaliar nos dados de teste)

Usei a métrica RMSE para avaliar a taxa de erro na estimativa dos preços de venda dos apartamentos.

Aplicar o Modelo **StackingCVRegressor (combinação de algoritmos)** a novos conjuntos de dados e realizar as previsões de preço de venda de apartamentos

Usar o resultado do modelo para suportar os usuários do Grupo ZAP na estimativa do melhor preço de venda do apartamento.

Deploy do modelo

Existem algumas formas de fazer o deploy de um modelo de Machine Learning em produção, geralmente categorizados em “Static” ou “Dynamic” e “On-demand” ou “Batch”

No caso desse projeto acredito que o aprendizado ocorre em modo offline, ou seja, o modelo é treinado uma vez em dados históricos. Se o modelo se tornar instável, será necessário reestruturá-lo.

Este projeto também possui uma caracteristica de obter previsões sob demanda, ou seja, previsões sendo realizadas em tempo real, no caso da estimativa de preços de apartamentos nos aplicativos ou web do Grupo ZAP.

Sendo assim, vejo que a melhor forma de publicação desse modelo de Machine Learning em produção seja através de Web Services, usando REST API. Um bom exemplo seria utilizar a biblioteca Openscoring.

Outros frameworks que trabalham com gerenciamento de workflow de modelos de Machine Learning são : Microsoft Azure Machine Learning e IBM com Watson (porém, ambos sendo necessário licença de uso)

Respostas da Entrega

Métrica: conforme apresentado no slide 7, a métrica RMSE é bastante aplicada em algoritmo de regressão e possui simplicidade na análise

Performance por faixas de preço e bairros: conforme apresentado no slide 10, utilizei o método de clusterização para segregar o dataset em 4 cluster a partir dos dados de bairro e preço.

3 campos para estimar: com base no algoritmo de Feature Selection (ExtraTreesRegressor), utilizaria os campos: neighborhood, monthlyCondoFee, bathrooms

Como você vislumbra colocar a sua solução em produção?

Apresentado no slide 12