Valuation Systems

Apresentação de case

Etapas da apresentação

- Contexto
- Dados e metodologia de tratamento.
- Construção do modelo
- Resultados
- Avaliação de risco
- Extrapolação do modelo

Contexto

Importância de avaliar corretamente o valor de um imóvel.

- Um mercado mais justo para o vendedor e comprador do imóvel
- O preço de um imóvel influencia diretamente o tempo para venda.
- Gerar estratégias para compras e vendas de imóveis.

Vantagens no uso de Machine learning para avaliação do preços de imóveis.

- Escalabilidade do processo de avaliação.
- Análises objetivas de preço.
- Redução no custo para avaliações acuradas.

Dados e metodologia de tratamento

Dados fornecidos

1. Localização (Rua, coordenadas e nome do edifício).
2. Número de quartos.
3. Número de garagens.
4. Área do imóvel.
5. Tipo de imóvel (Cobertura, Apartamento ou conjunto).
6. Valor do imóvel. (Valor padrão, máximo e mínimo) Output

Hipóteses iniciais sobre os dados

- Hipóteses sobre variáveis:
  - Características do imóvel. (Número de quartos, garagens e área do imóvel)
  - Regiões.
- Hipótese inicial sobre confiabilidade dos dados:
  - Dado mais confiável a localização via coordenadas.
  - Dado menos confiável nome da rua e nome do prédio.
  - Dados sobre quartos, área de imóvel e garagens são confiáveis.

Enriquecimento.

- Google Maps API. (Reverse geocoding das coordenadas)

  Dados obtidos

  1. Endereço completo (Rua, número, cep)
  2. Bairro
  3. Cidade
  4. País

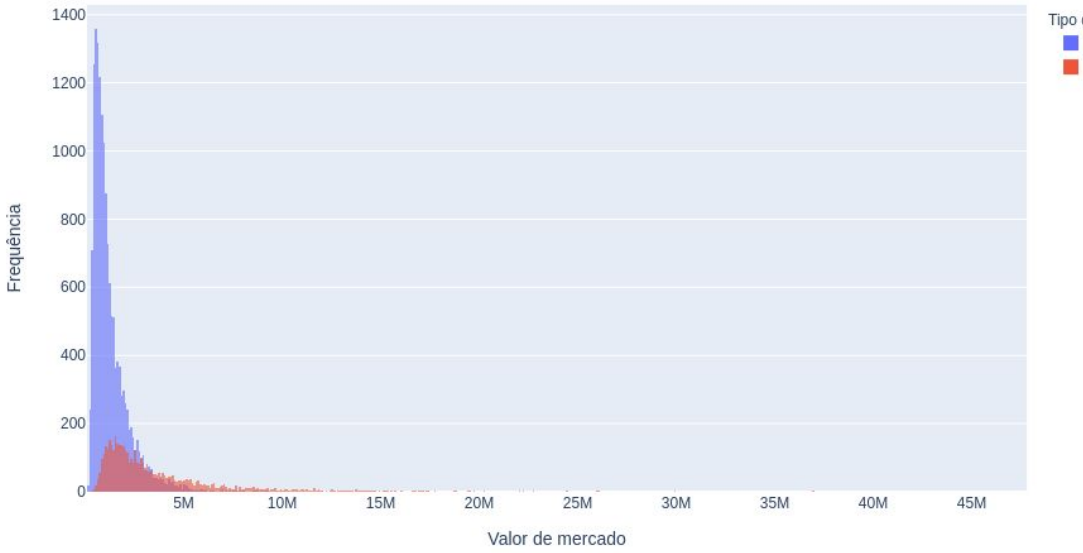
- Dados fora de São Paulo foram removidos. (285)

Necessário para a avaliação do modelo

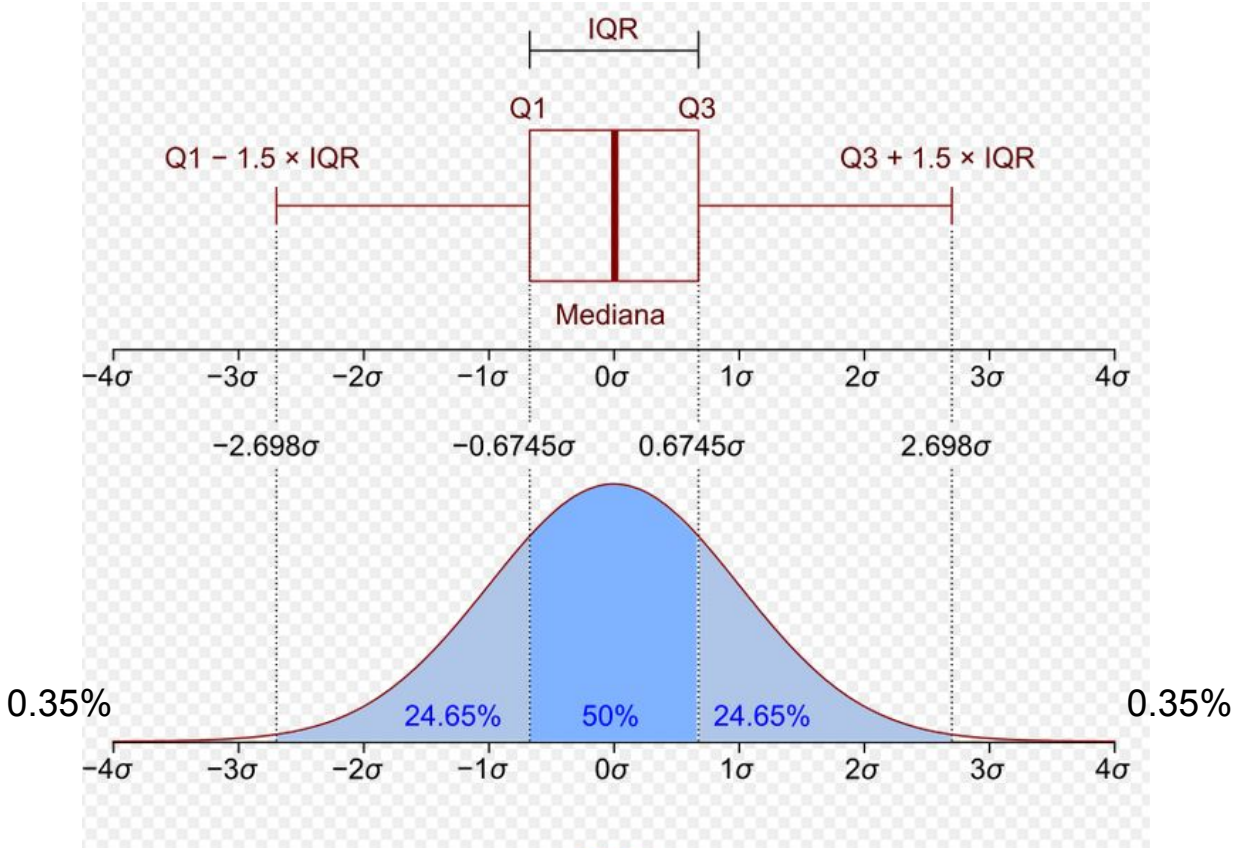
1. Features (Nº quartos, Nº garagens, área , localização)
2. Output (Valor de mercado, Valor máximo, Valor mínimo)

- Remoção de dados que possuem Nº quartos, Nº garagens e área faltando simultaneamente. (2935)
- Remoção de dados que não possuam output. (4043+1279)

Preço por tipo de imóvel.



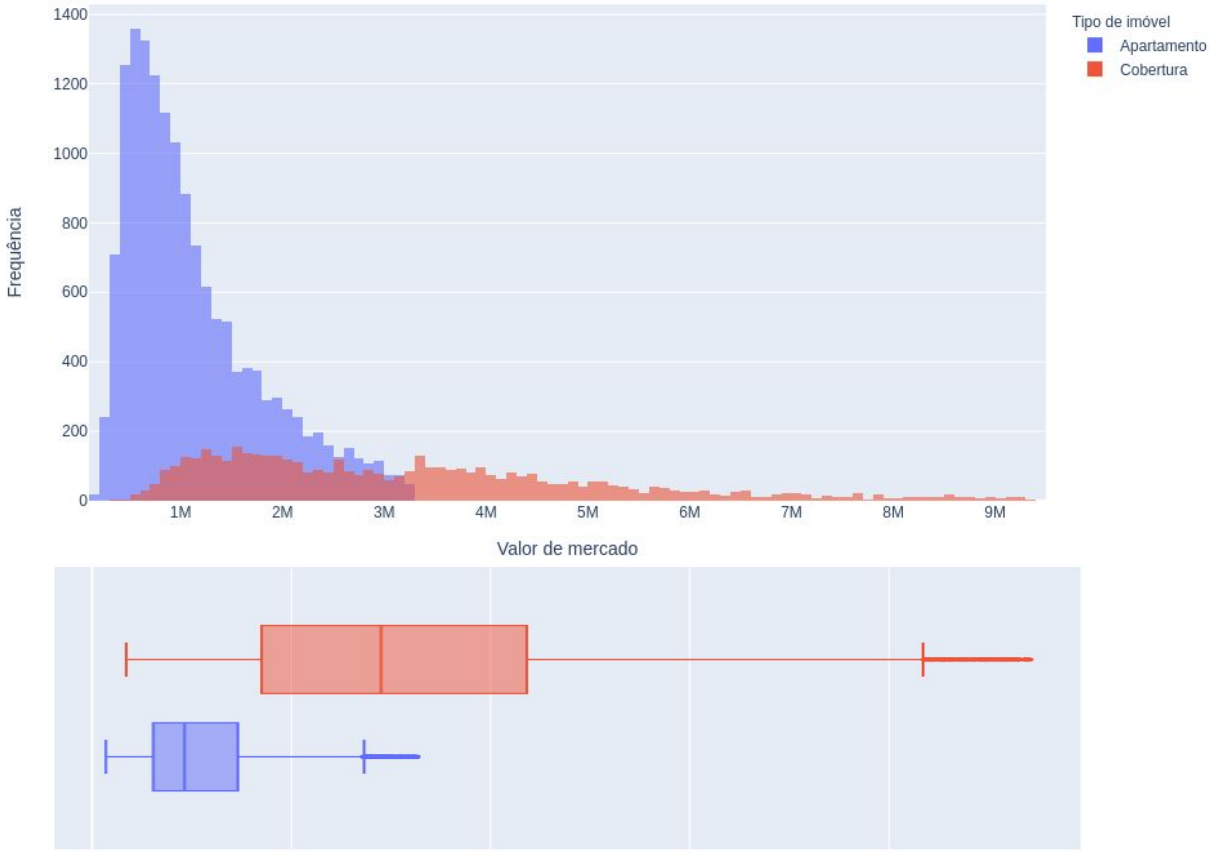
Descobrindo outliers.



Ações possíveis de serem tomadas

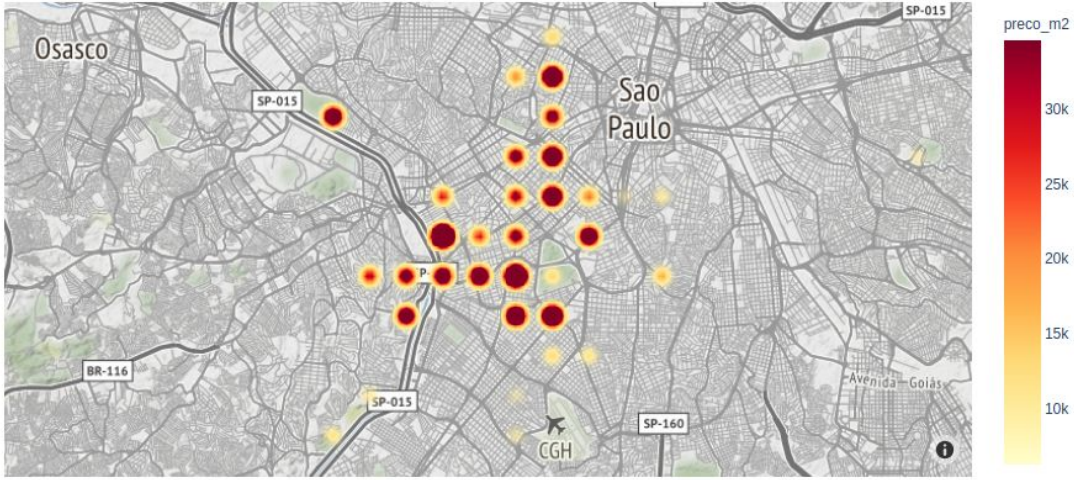
- Não remover os outliers
  - Valor máximo: $47.735.900. Nº de dados perdidos: 0
- Remover todos os outliers sem segmentação por tipo de imóvel
  - Valor máximo: $4.222.040 Nº de dados perdidos: 1707
- Remover os outliers de apartamento e cobertura:
  - Valor máximo: $7.850.110 Nº de dados perdidos: 1494
- Transformar os outliers de apartamento no tipo cobertura e remover os outliers de cobertura: (Utilizada)
  - Valor máximo $9.410.230. Nº de dados perdidos: 361
- Transformar os outliers de preço por metro quadrado de apartamento no tipo cobertura e remover os outliers de cobertura:
  - Valor máximo $36.947.700. Nº de dados perdidos: 187

Após remoção dos outliers



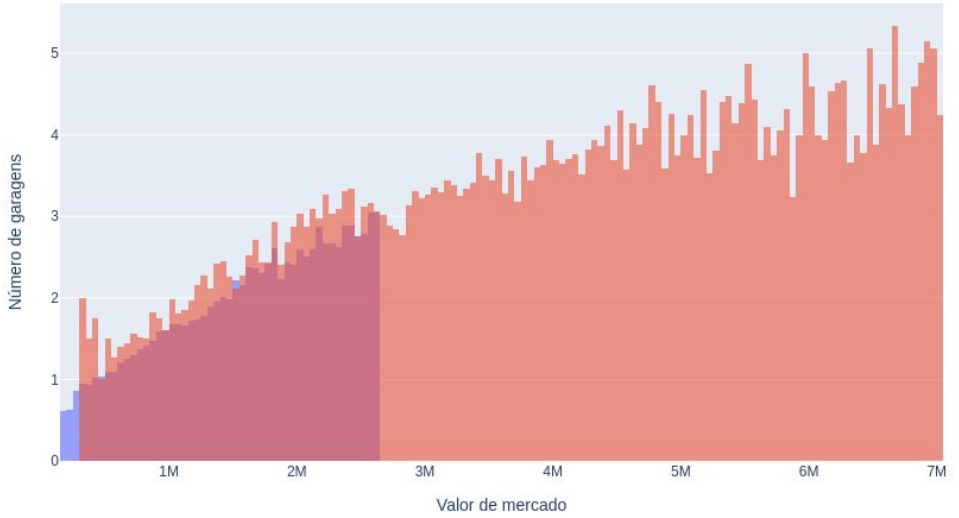
Outliers removidos

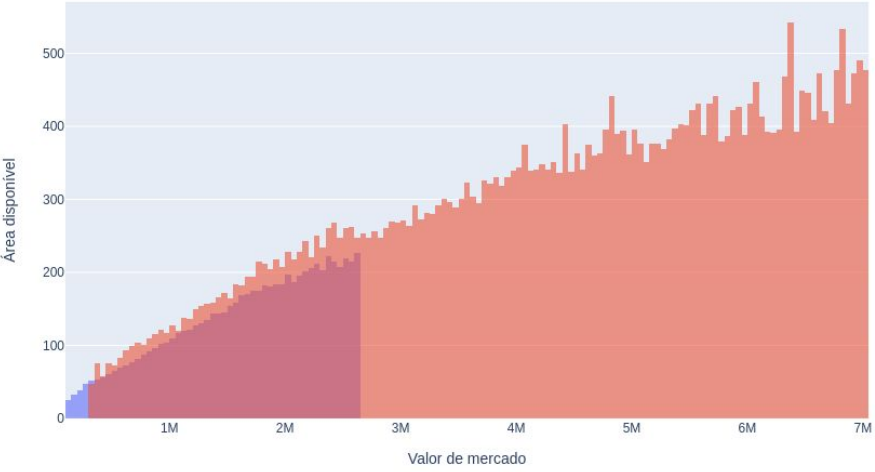
Retiro Morumbi terá previsões de baixa qualidade, já que só sobra mais um apartamento, após a remoção dos outliers.

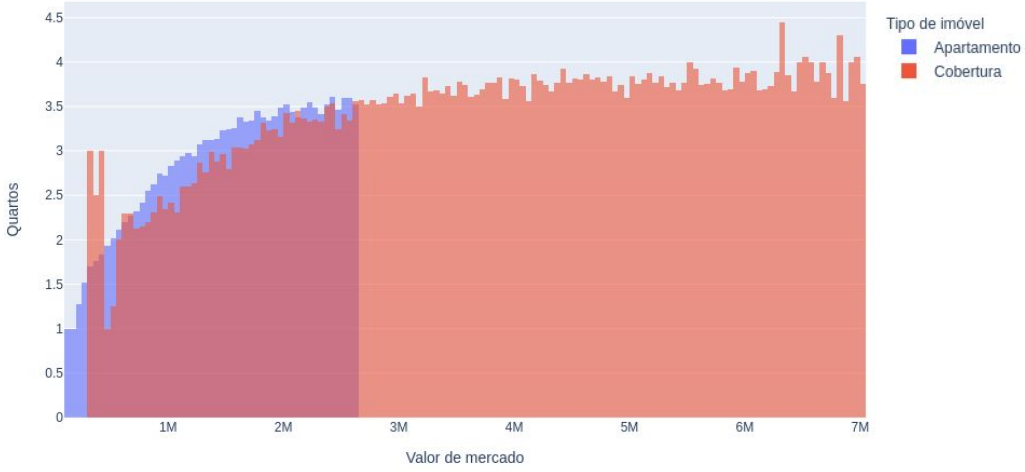


	Número de imóveis removidos	Número de imóveis restantes
Jardim Paulistano	65	320
Itaim Bibi	57	387
Cerqueira César	36	1026
Jardim Paulista	31	1177
Vila Buarque	30	610
Moema	26	297
Vila Nova Conceição	19	510
Alto de Pinheiros	17	112
Paraíso	17	524
Jardim Panorama	12	17
Higienópolis	9	325
Jardim Guedala	8	140
Bela Vista	5	873
Morumbi	5	16
Pacaembu	4	366
Pinheiros	3	786
Paraíso do Morumbi	3	111
Retiro Morumbi	2	1
Indianópolis	2	448
Campo Belo	2	312
Vila Mariana	2	700
Barra Funda	2	264
Aclimação	1	380
Vila Invernada	1	43
Jardim Ampliacao	1	149
Santana	1	286

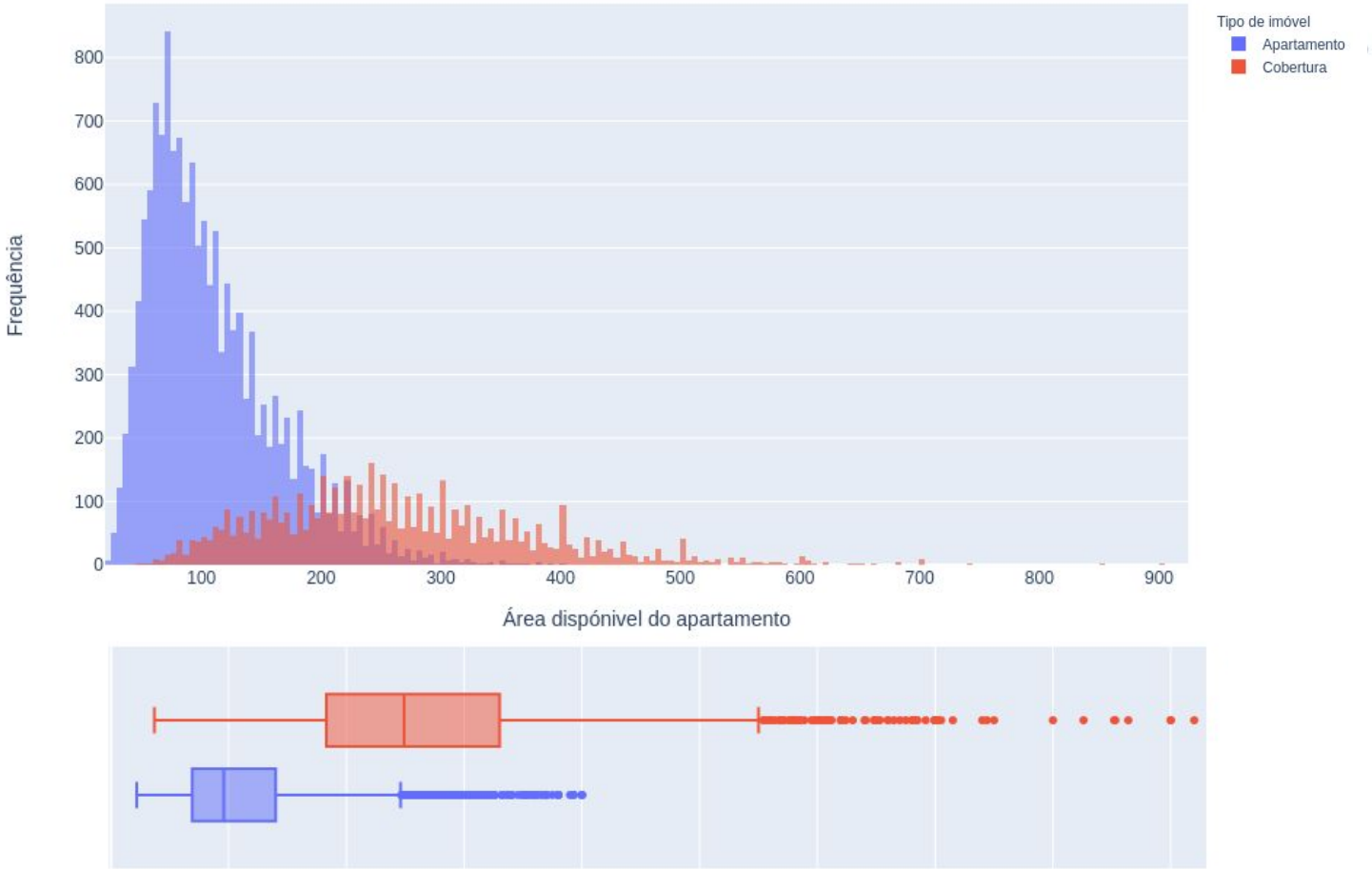
Variáveis quantitativas.



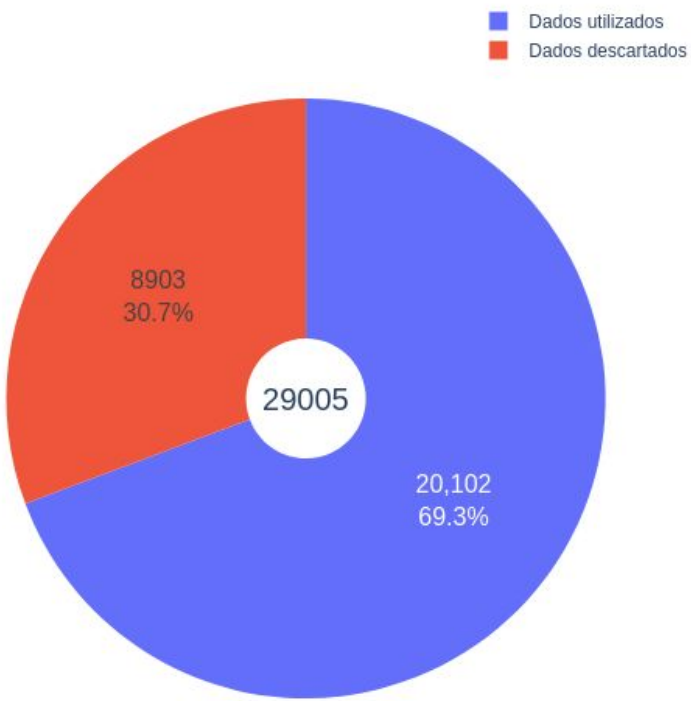


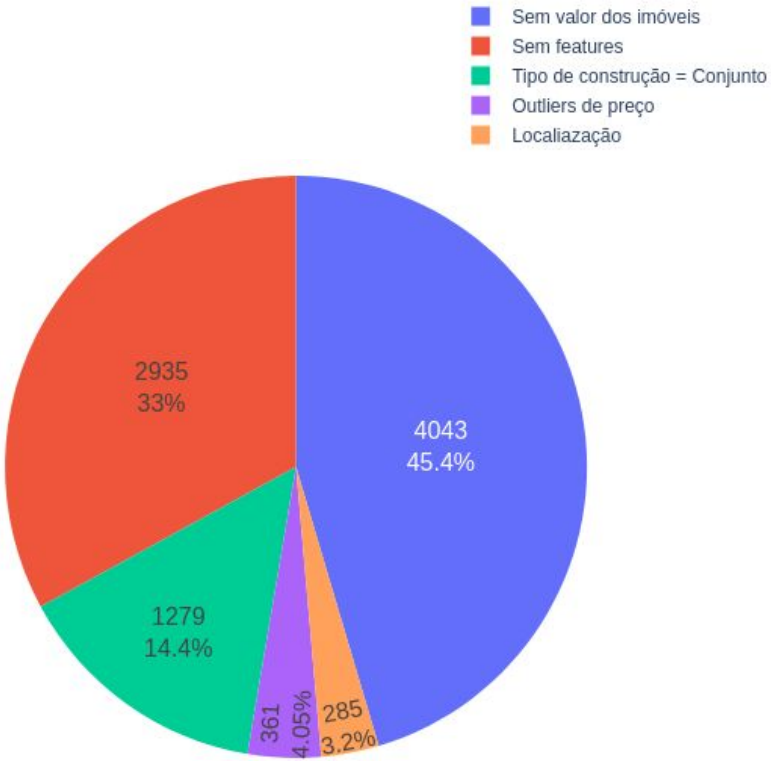


Outros outliers



Uso de dados


Dados utilizados
Dados descartados
8903
30.7%
29005
20,102
69.3%


Sem valor dos imóveis
Sem features
Tipo de construção = Conjunto
Outliers de preço
Localiazação
2935
33%
4043
45.4%
1279
14.4%
361
4.05%
285
3.2%

Imputação de dados.

Após a remoção dos dados foi necessário realizar a imputação de dados de garagens e de três apartamentos.

Levando em consideração a redução no número de garagens nos últimos anos graças ao plano diretor de São Paulo, junto com ser uma estratégia que melhorou a performance do modelo, quando comparada com outras, é assumido que todos os valores de garagens faltando é zero.

Sobre os 3 apartamentos que ficaram sem dados sobre o número de quartos, esses dados foram encontrados após uma pesquisa, utilizando os dados de localização do apartamento e observando se o número de garagens e tamanho do apartamento batiam.

Limitações por remoção de dados

- Previsão de valores maiores que 9.4M.
- Previsão de imóveis do tipo Conjunto

Construção do modelo

Novas features

- Bairros.
  - Todos os bairros com pelo menos 1% do total de apartamentos do dataset de treino.
  - Bairros com menos de 1% foram agrupados na classe “Outro”.
  - De 165 bairros se reduz a 36

- Valor dos bairros.
  - Obter a média do preço por metro quadrado dos bairros, no dataset de treino.
  - Separar em 10 grupos de tamanhos iguais, por faixa de preço.

:

	preço m2
0	(4838.491, 5436.672]
1	(5436.672, 6524.445]
2	(6524.445, 7107.352]
3	(7107.352, 7511.266]
4	(7511.266, 8114.268]
5	(8114.268, 8587.035]
6	(8587.035, 9457.777]
7	(9457.777, 10172.537]
8	(10172.537, 11496.573]
9	(11496.573, 15692.824]

Modelos

- XGBoost
  1. Algoritmo baseado em árvores de decisões.
  2. Ótimo para realizar previsões de dados não estruturados.
  3. Ruim para fazer extrapolações.
  4. Realiza boosting
- Random Forest
  1. Algoritmo baseado em árvores de decisões
  2. Benchmark para o XGBoost.
  3. Ruim para fazer extrapolações

Workflow avaliação do modelo.

Enriquecimento de dados → Limpeza de dados → Remoção de outliers

Cross validation externo → Treino teste → Feature engineering (Treino)

Cross validation interno → Tunning de hiperparâmetros → Volta cross validation externo

Treina o modelo → Obtém a performance → Realiza a média sobre métricas obtidas

Métricas de avaliação

1. Root Mean Squared Error (RMSQ):
   - Avaliação de erros sobre outliers.

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

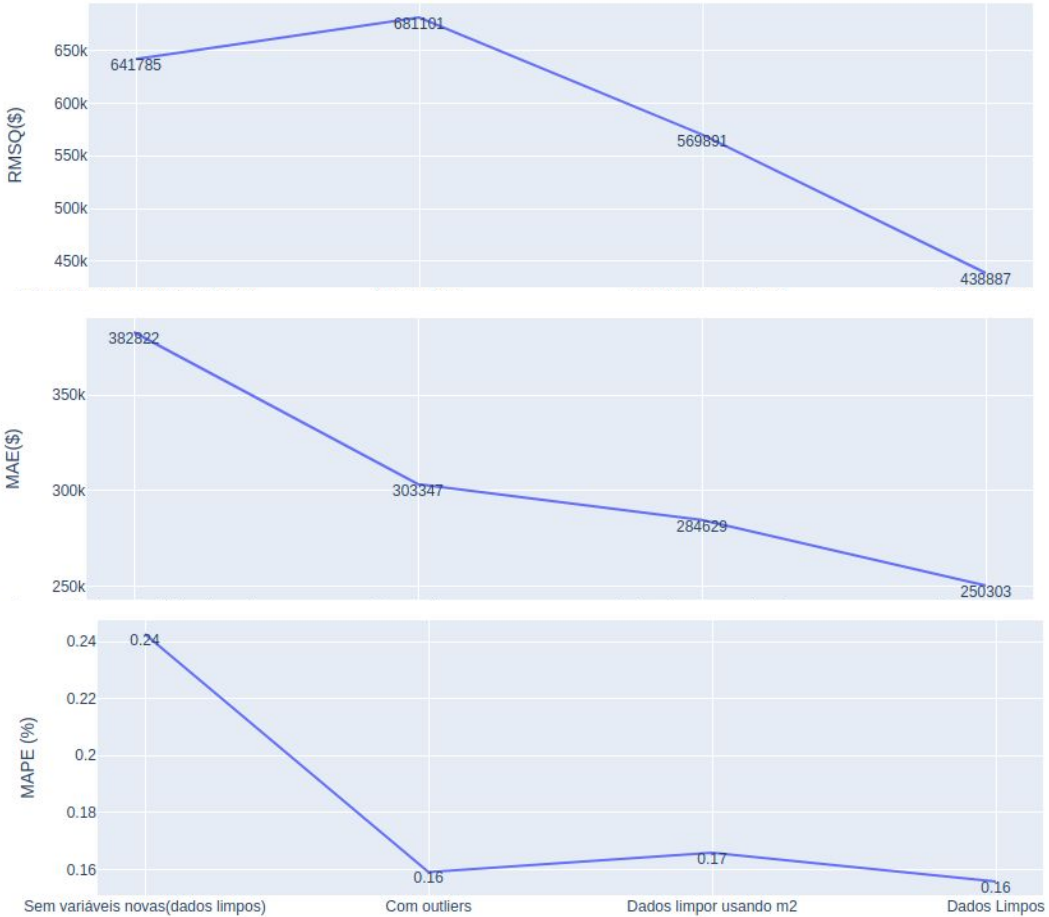
2. Mean absolute error (MAE):
   - Avaliação geral de erros.

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

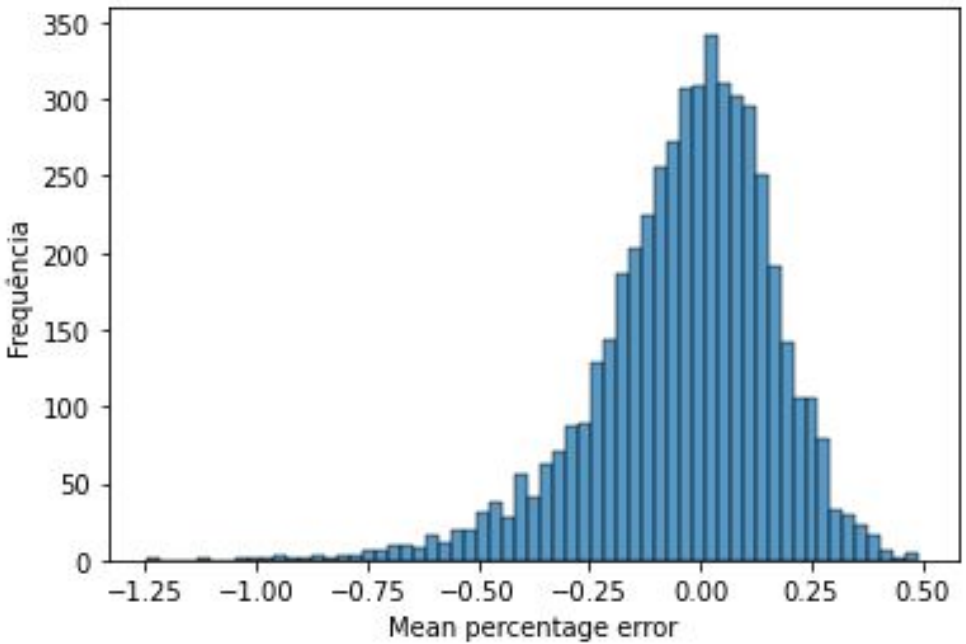
Resultados

XGBoost X Random Forest

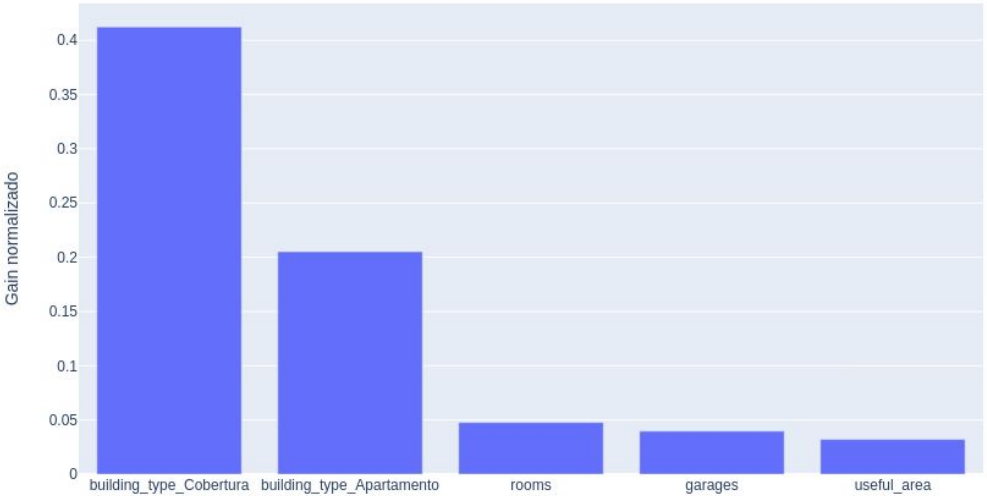
Resultado das principais métricas

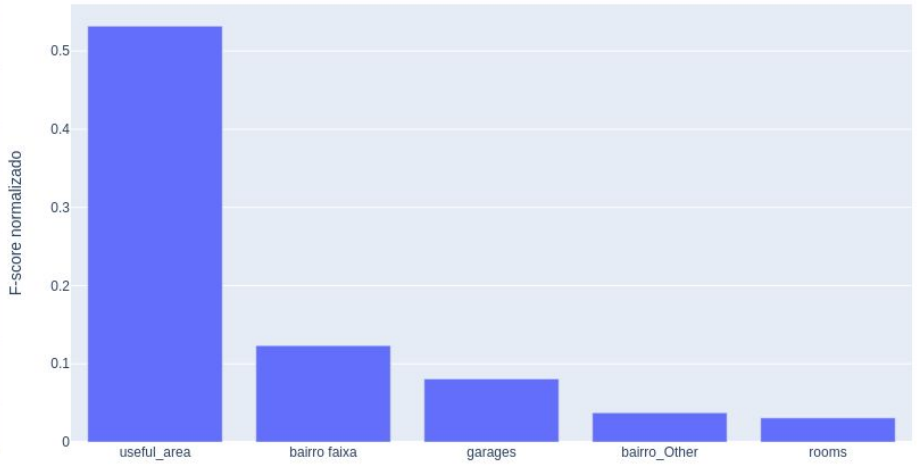


Distribuição de erros do XGBoost



Importância das features

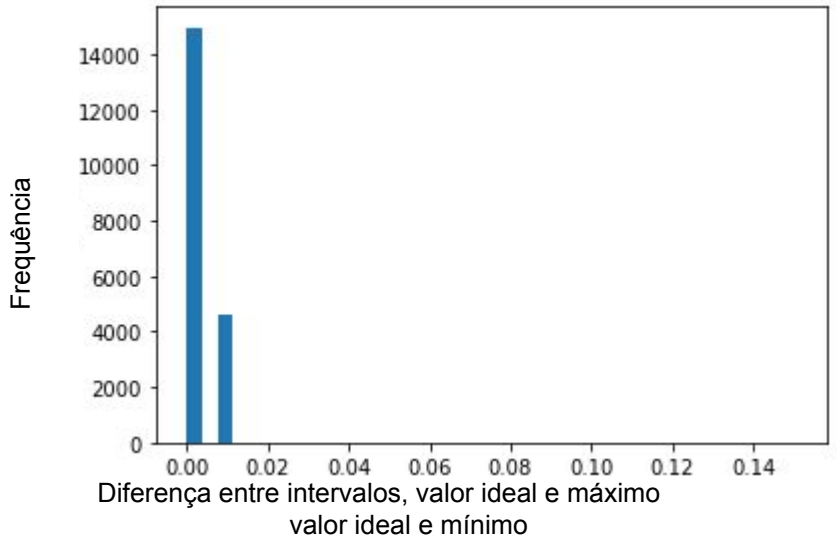




Avaliação de risco

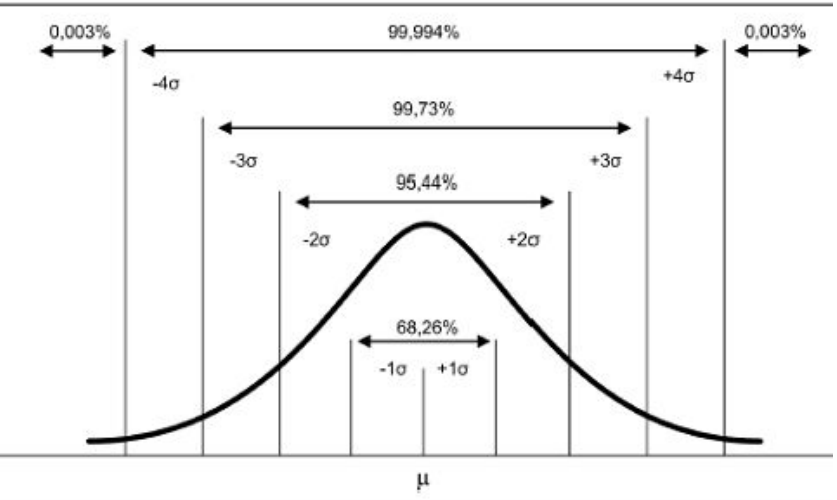
Preço desejado x Preço de mercado

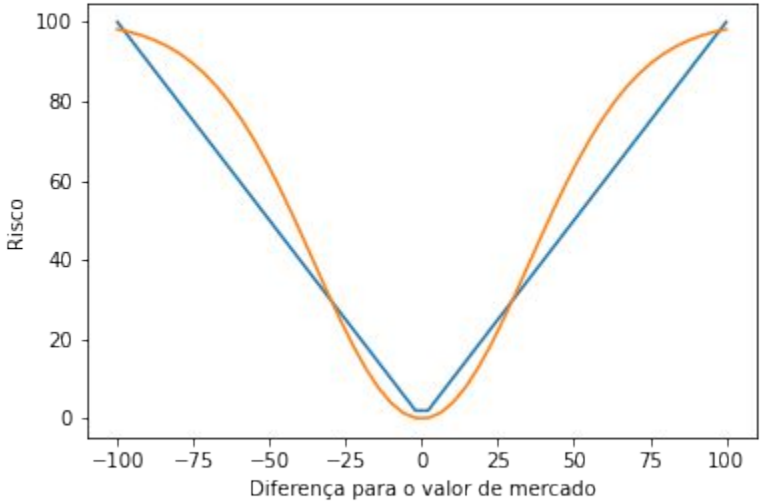
- Como avaliar se o preço desejado condiz com os valores de mercado?
- São realizadas previsões para o valor mínimo, máximo e ideal.
- Risco para o valor ideal = 0, para valores máximos e mínimos = 100.
- Obs: Intervalos entre o ponto ideal e os valores máximos e mínimo é aproximadamente igual.



Formas de medir o risco

- Distribuição linear possui um decaimento constante.
- Distribuição normal possui decaimento lento próximo da média e mais rápido no final.



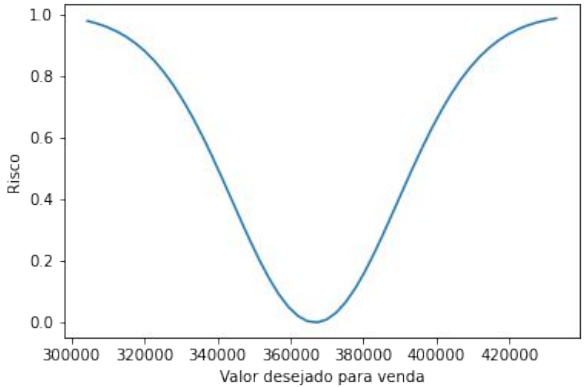


Preço desejado x Realidade do mercado

1. Média da normal = Valor médio das previsões de valor máximo, mínimo e estimação pontual.
2. Desvio padrão = (Valor máximo-Valor mínimo)/4

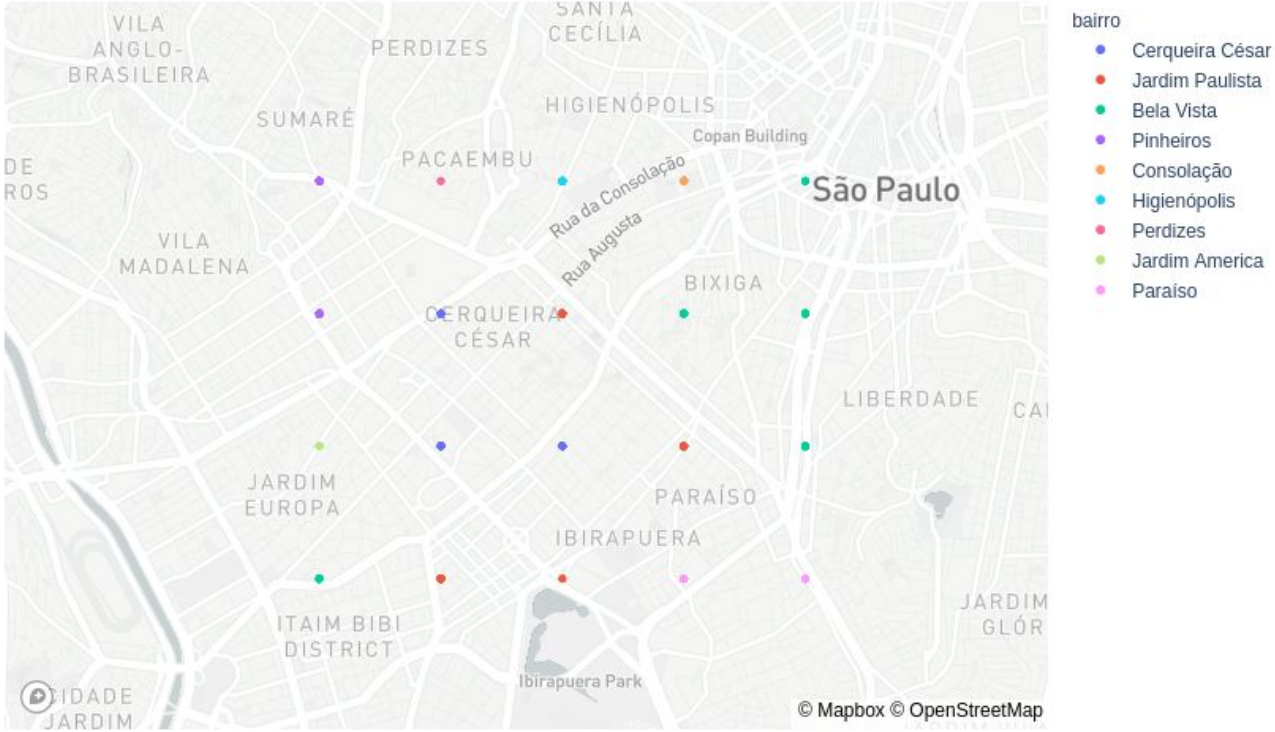
$$f(x) = e^{\frac{-(x-\mu)^2}{\sigma^2}}$$

Ex: Estimação pontual = 366753, valor máximo = 432769, valor mínimo = 304405

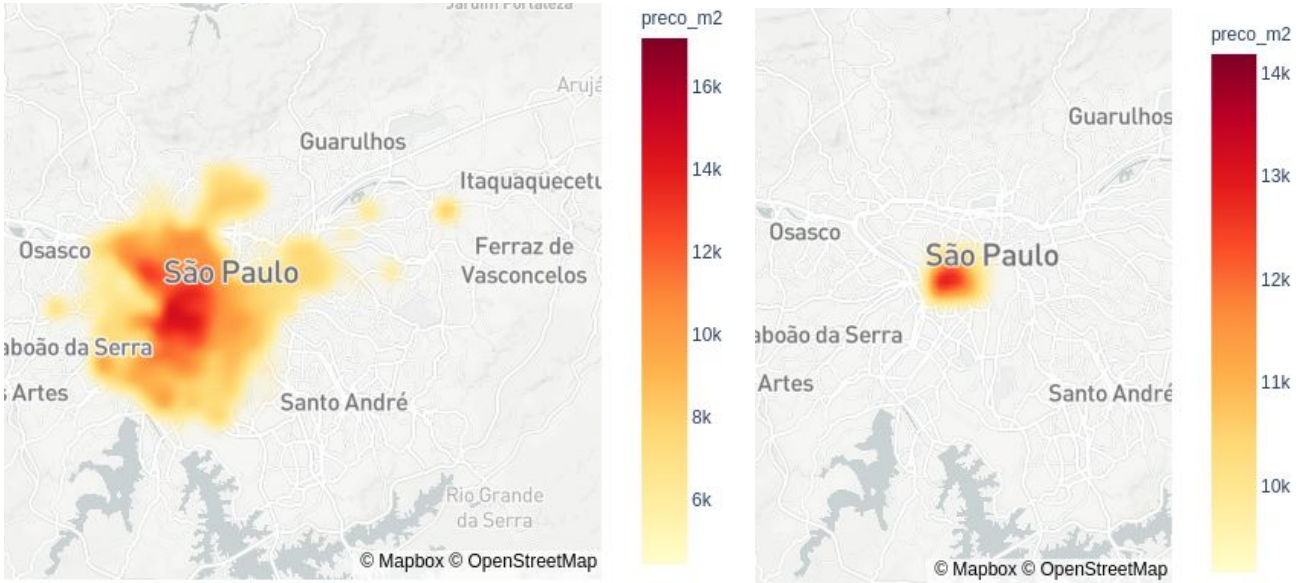


Extrapolação do modelo

Bairros do bounding box



Extrapolação do modelo



Extrapolação do modelo

1. As variáveis de localização funcionam no bounding box. Mas fora elas não conseguem dar informação.
2. Nesse caso o modelo realizaria previsões baseado somente no N de garagens, N de quartos e tamanho do apartamento.
3. O modelo funcionará bem em regiões com preço por metro quadrado semelhantes ao do bounding box.
4. No caso da região o modelo funciona bem em outras regiões que possuem um alto valor do preço por metro quadrado.

MAE e RMSE para o caso da extrapolação do modelo

Escolha de 1000 dados

- Dados entre o primeiro e o terceiro quartil do valor estimado.
- Excluiria outliers de número de quartos, garagens e tamanho da casa.
- Dos dados restantes, tentaria manter um equilíbrio entre diversidade de bairros e quantidade de amostras.

Eficiência do modelo por tamanho de treino

- O modelo possui uma performance cerca de 20% menor com 500 dados, e 15% menor com 1000 dados.