

Case para Seleção



Autor:
José Grimaldo

Sumário

- Introdução
- Objetivos
- Metodologia
- Método
- Resultados
- Considerações finais

Introdução

Case para seleção do






- Bolsista Programa Inova Talentos - IEL / Cnpq
- Duração de ~8 meses

Foco em aprendizagem de máquina

- Aplicação prática
- Análise de resultados

Objetivo do case

Desenvolver uma solução para

- Verificar a qualidade dados 
- Encontrar informações sobre as variáveis 
- Apresentar variáveis mais relevantes dessa base 
- Apresentar as oportunidades de negócio 
- Apresentar métricas de erro do modelo 

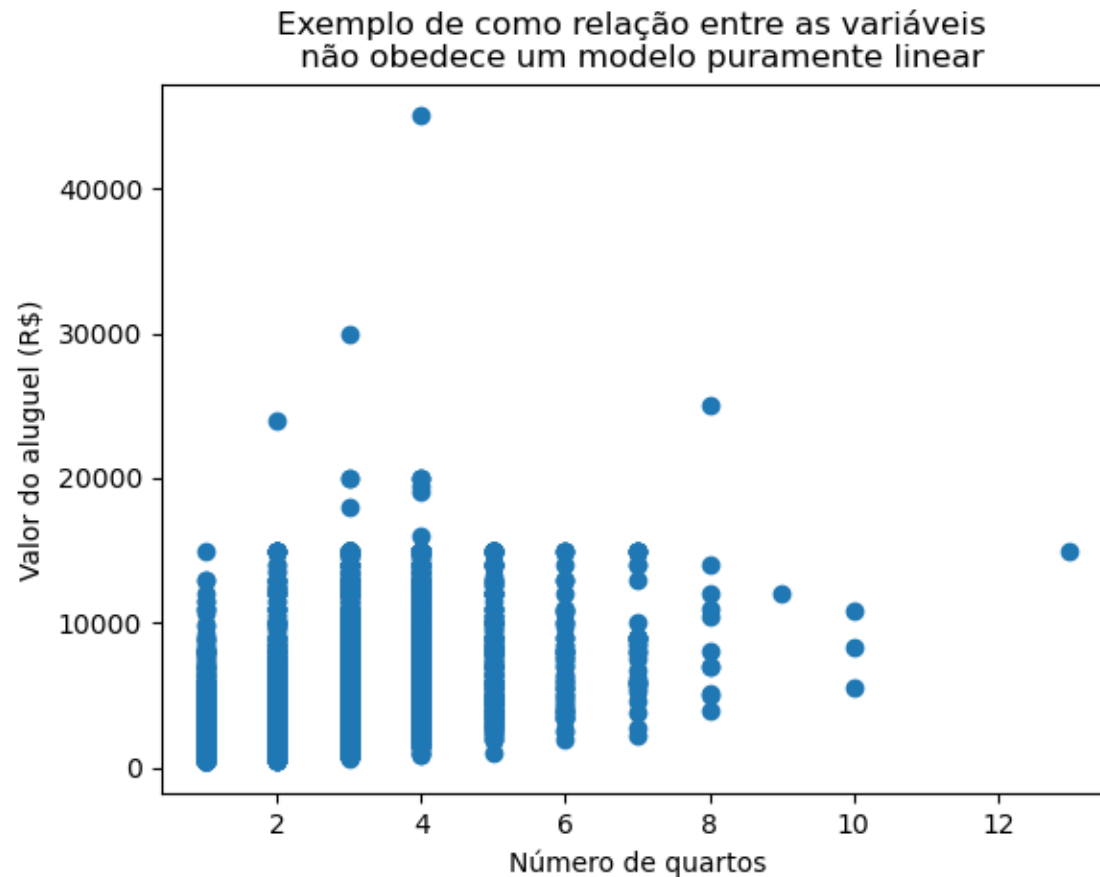
Metodologia

- Analisar os dados e tipo de predição desejada
- Observar tipos de método compatíveis
- Análise de bibliotecas compatíveis
- Separar dataset em treino (85%) e teste (15%)
- Treinar método utilizando o subset de treino
- Análise utilizando R^2 no subset de teste

Informações gerais sobre variáveis

Relação entre preditor e o preço do aluguel não é linear

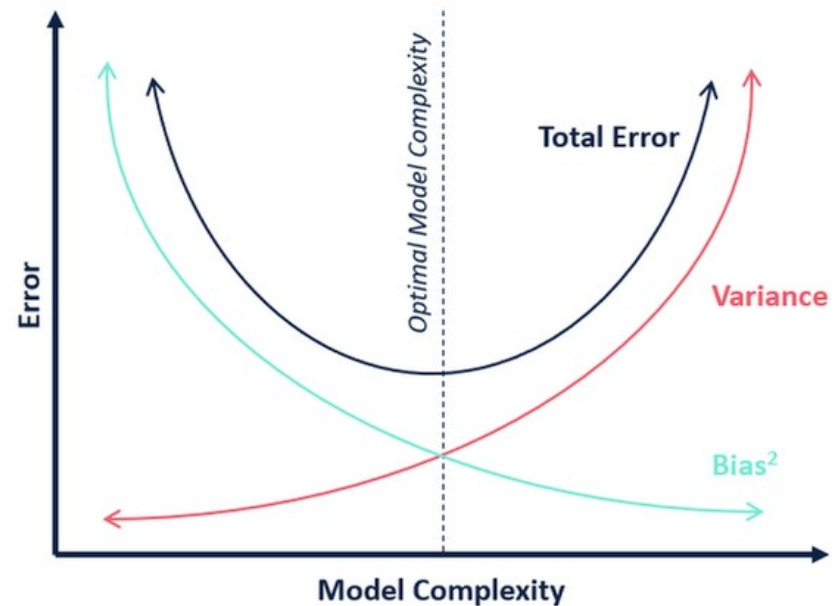
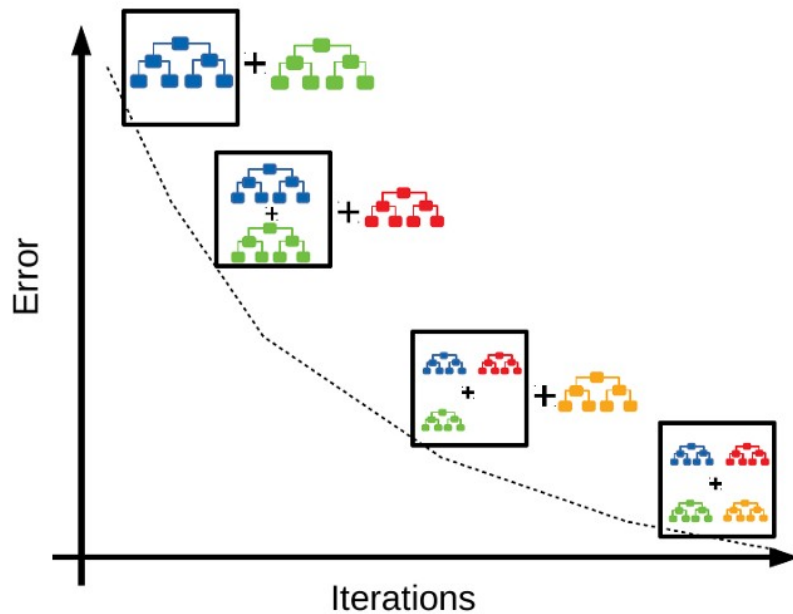
- Além disso existem “*outliers*”



Método

Gradient Boosting

- Regressão Linear é insuficiente
- Gradient boosting satisfaz os critérios
- Similar a Random Forests



Treinamento

Treinamento e calibração do boosting

- Tempo para calibração ~5 minutos
- Tempo de treinamento 1.82s
- Tempo de predição $<0.01s$

Para treinar dados foram transformados em formato numérico

- Nome de cidade
- Mobiliado (Sim/Não)
- Aceita animais (Sim/Não)
- Andar (“-” ou um número)

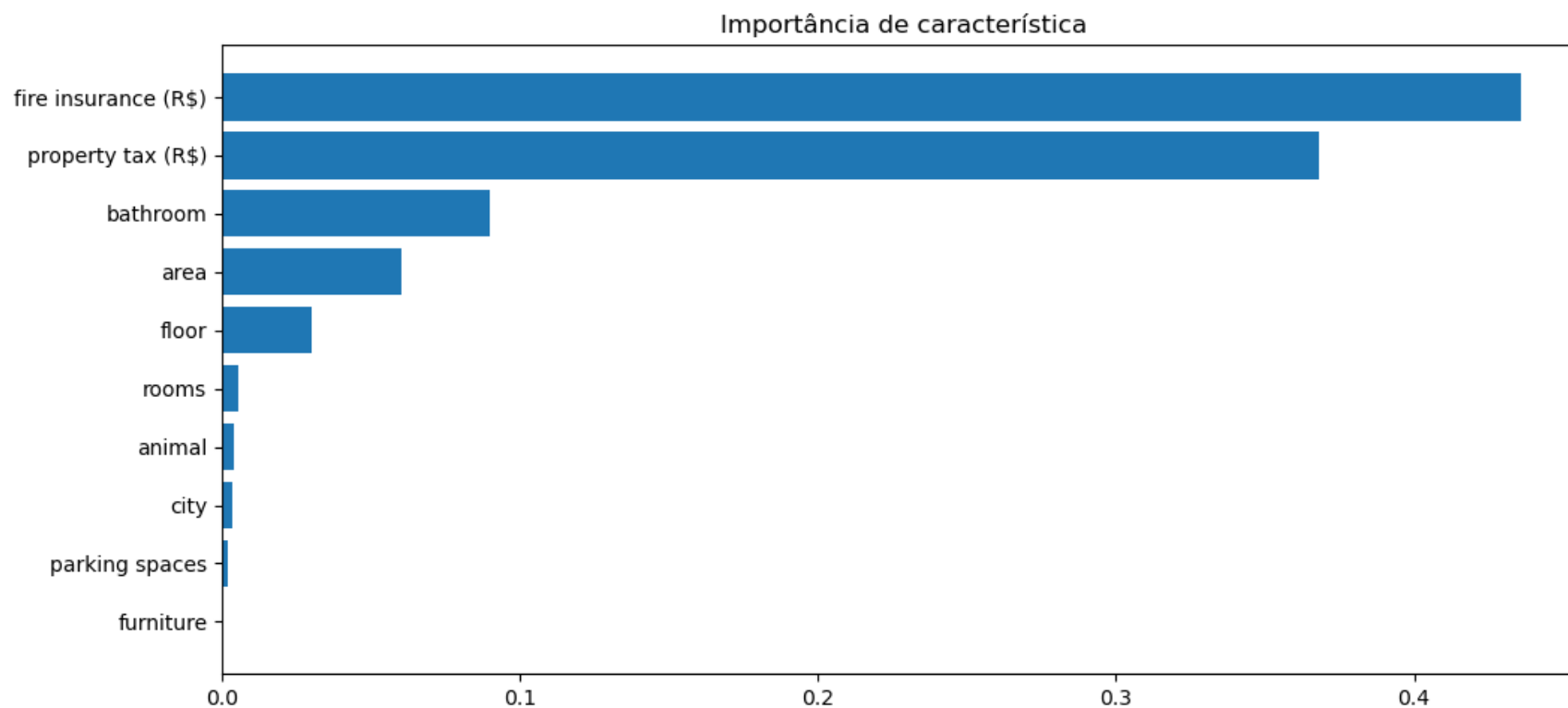
Resultados

Análise de resultados foi feita com R^2

- Análise realizada no conjunto teste
- R^2 foi de 0.90 (ideal seria 1.0)
- Fórmula:

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

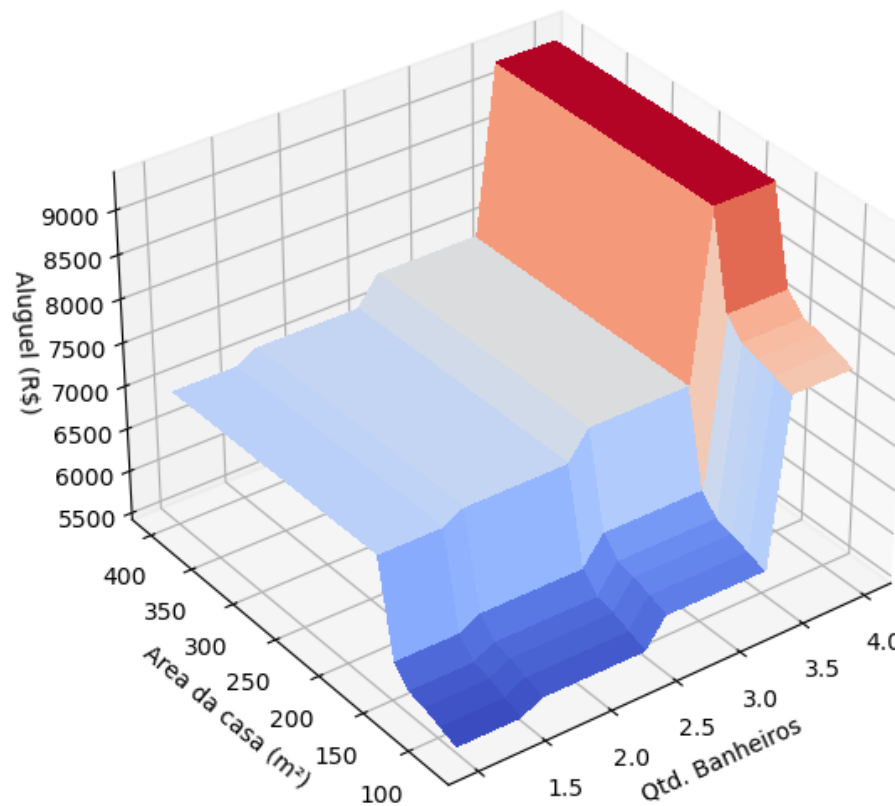
Características importantes



Decisões de Negócio

O treinamento do modelo permite a realização de inferências diversas, por exemplo

Relação entre qtd de banheiros, area de casa e o impacto no aluguel



Considerações Finais

O método treinado no dataset permite interpolar o custo do aluguel baseado em preditores

- Baixo número de preditores facilita análise
- Ausência de datas dificulta a extrapolação
- Aumentar dimensionalidade poderia motivar o uso de Deep Learning (NN)
- Impacto da “Curse of dimensionality” é um fator