CUSTO DAS HABITAÇÕES DA CALIFÓRNIA

OBJETIVO PRINCIPAL

Aprendizagem de máquina da base de dados

OBJETIVOS ESPECÍFICOS

Modelar os dados

Identificar o modelo mais eficiente nas previsões

Métodos

FASE 1

Análise descritiva e pré-processamento dos dados

FASE 2

Regressão linear simples e múltipla

FASE 3

Validação Cruzada simples e o método k-fold

Modelos de regressão

1. Ajuste dos modelos
   Análise residual
   Análise de diagnóstico

2. Multicolinearidade
   Transformação de Box-cox

Validação Cruzada

1. Método simples: divisão dos dados em conjuntos de treinamento e teste
   Métodos k-fold: divisão dos dados em k subconjuntos

2. Análise de medidas: R-quadrado, RMSE (erro médio quadrático da raiz) e MAE (erro médio absoluto)

Análise descritiva

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
1	-122.23	37.88	41	880	129	322	126	8.3252	452600	NEAR BAY
2	-122.22	37.86	21	7099	1106	2401	1138	8.3014	358500	NEAR BAY
3	-122.24	37.85	52	1467	190	496	177	7.2574	352100	NEAR BAY
4	-122.25	37.85	52	1274	235	558	219	5.6431	341300	NEAR BAY
5	-122.25	37.85	52	1627	280	565	259	3.8462	342200	NEAR BAY
6	-122.25	37.85	52	919	213	413	193	4.0368	269700	NEAR BAY
7	-122.25	37.84	52	2535	489	1094	514	3.6591	299200	NEAR BAY
8	-122.25	37.84	52	3104	687	1157	647	3.1200	241400	NEAR BAY
9	-122.26	37.84	42	2555	665	1206	595	2.0804	226700	NEAR BAY
10	-122.25	37.84	52	3549	707	1551	714	3.6912	261100	NEAR BAY

Em média, as casas possuem aproximadamente 29 anos, localizadas a -119,6 de longitude e 35,63 de latitude, possuindo 2.636 cômodos, incluindo cerca de 538 quartos e custando 206.856 dólares. Cada grupo de casas é formado, em média, por 1425 pessoas, sendo cerca de 500 famílias com renda média de 3,87 dezenas de milhares dólares.

Pré-processamento

- 207 informações faltantes no atributo da quantidade de quartos substituídos pela média do atributo
- Normalização de todos os atributos

Regressão linear simples

- Modelo com variável resposta valor médio das casas e variável independente renda média (maior correlação)
- Existe relação linear entre as variáveis
- O parâmetro (beta 1) da variável renda média está bem estimado para a reta

Regressão linear múltipla

- Modelo com variável resposta valor médio das casas e todas as demais variáveis numéricas como variáveis independentes
- O Teste T rejeitou a hipótese de que todos os coeficiente são diferentes de zero
- O teste F, afirmou que existe pelo menos uma variável que mantém relação linear com a variável resposta
- R-quadrado com um bom ajuste (63,56%)
- Nenhum pressuposto estava sendo atendido (normalidade, homocedasticidade, linearidade, autocorrelação)
- Retirada dos outliers (pontos aberrantes que não são de alavanca nem de influência) - pouca mudança na análise de resíduos

Multicolinariedade

- Condição que ocorre quando algumas variáveis preditoras no modelo estão correlacionadas a outras variáveis preditoras
- Uma multicolinearidade forte é problemática porque pode aumentar a variância dos coeficientes de regressão, tornando-os instáveis
- Uso do VIF (fatores de inflação da variância): mede o quanto a variância de um coeficiente de regressão estimado aumenta se seus preditores estão correlacionado
- Variáveis com VIF > 2 eliminadas, restando o valor médio das casas, idade das casas e renda média.
- Queda do R-quadrado

Transformação de Box-cox

- Tranformação da variável resposta: transformação de Box-cox
- Uma das possíveis formas de contornar o problema de dados que não obedecem os pressupostos da análise de resíduos
- Nova análise de resíduos: pressuposto de normalidade atendido

Validação Cruzada

Tabela 4: Validação cruzada - método simplificado

R-quadrado	RMSE	MAE	Modelo
0,4770629	82582,82	61797,67	MRLS
0,6368851	68813,88	50165,35	MRLM
0,509299	80030,74	59544,75	MRLM com multicolinariedade
0,5092593	236337,1	206924	MRLM com boxcox

Tabela 5: Validação cruzada - método k-fold

R-quadrado	RMSE	MAE	Modelo
0,472938	84015,19	62883,02	MRLS
0,6337874	70022,15	51180,54	MRLM
0,5194356	78061,2	58090,89	MRLM com multicolinariedade
0,4905648	15,07849	11,69102	MRLM com boxcox

Conclusão

- A Regressão Linear Múltipla se ajustou melhor que a Linear Simples
- O melhor ajuste da Regressão Linear Múltipla foi com a retirada de algum dos outliers, com a nova seleção de variáveis com multicolinariedade e com a transformação de Box-cox