House Pricing

Hirley Dayan Louren?o da Silva e Marcia Maria Parmigiani Martins

Leitura dos Dados

O dataset utilizado nesse trabalho refere-se a dados de im?veis como total de c?modos, idade, etc, o qual foi dividido entre treino, valida??o e teste, com o objetivo de cria??o de um modelo utilizando algoritmo de **Regress?o Linear** para predi??o do pre?o de im?veis.

O dataset de treino possui **12384** observa??es e **10** features.
O dataset de valida??o possui **4128** observa??es e **10** features.
O dataset de teste possui **4128** observa??es e **10** features.

Tratamento dos dados

O dataset possui uma feature categ?rica **ocean_proximity** com 5 n?veis. Considerando que a Regress?o Linear assume que todas as vari?veis independentes s?o num?ricas, iremos utilizar a t?cnica de **hot encoding** para transformar a feature em num?rica, atribuindo valor 1 se o caso se enquadre na determinada categoria. A inclus?o da feature categ?rica possibilitou um resultado melhor para todos os modelos testados.

Al?m disso, foram removidas observa??es de features sem anota??es (NA) nos dados de treino, valida??o e teste.

Normaliza??o dos dados

A t?cnica de normaliza??o aplicada ao dataset ? a **Min-Max** onde os dados s?o dimensionados em um intervalo fixo, normalmente de 0 ? 1. A feature target **median_house_value** n?o foi inclu?da na normaliza??o.

longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income
0.61	0.16	1.00	0.05	0.07	0.05	0.07	0.13
0.60	0.15	0.88	0.03	0.05	0.02	0.05	0.18
0.20	0.58	0.63	0.00	0.00	0.00	0.00	0.25
0.53	0.30	0.67	0.00	0.00	0.01	0.01	0.10
0.57	0.17	0.63	0.14	0.12	0.06	0.12	0.43

Regress?o Linear

O algoritmo de **Regress?o Linear** foi utilizado para predizer os pre?os dos im?veis onde a vari?vel que se deseja encontrar ? a **median_house_value** que representa o valor do im?vel baseado em suas features.

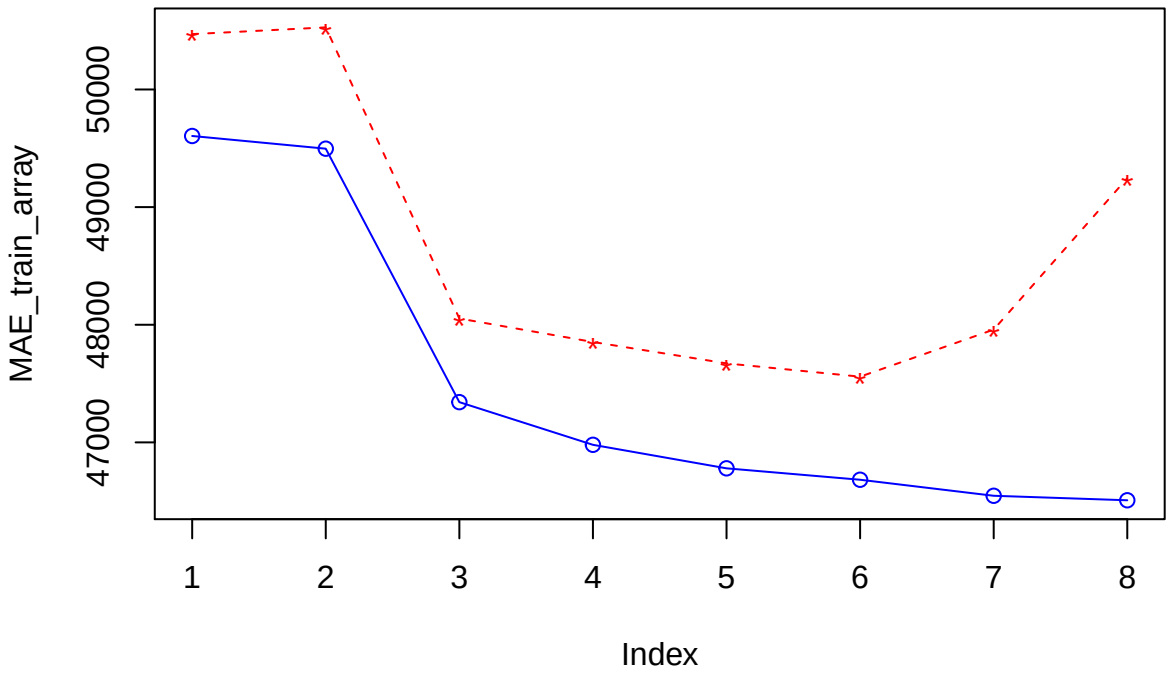
As medidas de avalia??o utilizadas foram a **MAE (Mean Absolute Error)** que calcula a m?dia da diferen?a absoluta entre os valores preditos e os observados e o Coeficiente de Determina??o (R2) que indica qu?o bem o modelo consegue se ajustar sobre um conjunto de predi??es e seus valores verdadeiros, podendo variar entre 0 e 1, sendo que 0 indica que o modelo n?o consegue explicar a variabilidade dos dados e 1 indica que as predi??es se ajustam perfeitamente aos dados (modelo explica toda a variabilidade das predi??es) o que indicar? **overfitting**.

Resultado dos modelos

Avalia??o para o modelo **baseline**: MAE **50470.57** e R? **63.43**%.

Avalia??o para o modelo **complexo baseado na combina??o de features existentes**: MAE **50999.99** e R? **63.43**%.

Avalia??o para os modelos **complexos baseado em regress?o polinomial**:
F?rmula grau 1: MAE = **50470.57** e R? = **63.43**%
F?rmula grau 2: MAE = **50527.82** e R? = **63.91**%
F?rmula grau 3: MAE = **48052.91** e R? = **66.26**%
F?rmula grau 4: MAE = **47853.44** e R? = **66.81**%
F?rmula grau 5: MAE = **47671.35** e R? = **67**%
F?rmula grau 6: MAE = **47558.52** e R? = **67.21**%
F?rmula grau 7: MAE = **47958.79** e R? = **61.92**%
F?rmula grau 8: MAE = **49245.76** e R? = **-19.87**%



A partir da fun??o de grau 7 percebe-se que os modelos tem um bom desempenho nos dados de treinamento, por?m n?o generaliza, tornando-se muito especializado, o que caracteriza **overfitting**.

Resultado do melhor modelo no conjunto de teste

O modelo escolhido para uso no conjunto de teste ? o polinomial de grau 6, pois na curva do MAE foi o que apresentou menor diferen?a entre o valor real e o predito durante o treinamento. A escolha da fun??o foi feita de modo a garantir uma converg?ncia entre o treinamento e a valida??o.

Avalia??o para o modelo **polinomial de grau 6**: MAE **46984.26** e R? **68.44**%.