House Pricing

Hirley Dayan Lourenço da Silva e Marcia Maria Parmigiani Martins

## Leitura dos Dados

O dataset utilizado nesse trabalho refere-se a dados de imóveis como total de cômodos, idade, etc, o qual foi dividido entre treino, validação e teste, com o objetivo de criação de um modelo utilizando algoritmo de **Regressão Linear** para predição do preço de imóveis.

O dataset de treino possui **12384** observações e **10** features.  
O dataset de validação possui **4128** observações e **10** features.  
O dataset de teste possui **4128** observações e **10** features.

## Tratamento dos dados

O dataset possui uma feature categórica **ocean\_proximity** com 5 níveis. Considerando que a Regressão Linear assume que todas as variáveis independentes são numéricas, iremos utilizar a técnica de **hot encoding** para transformar a feature em numérica, atribuindo valor 1 se o caso se enquadre na determinada categoria. A inclusão da feature categórica possibilitou um resultado melhor para todos os modelos testados.

Além disso, foram removidas observações de features sem anotações (NA) nos dados de treino, validação e teste.

## Normalização dos dados

A técnica de normalização aplicada ao dataset é a **Min-Max** onde os dados são dimensionados em um intervalo fixo, normalmente de 0 à 1. A feature target **median\_house\_value** não foi incluída na normalização.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| longitude | latitude | housing\_median\_age | total\_rooms | total\_bedrooms | population | households | median\_income | lt1\_h\_ocean | inland | near\_bay | near\_ocean | median\_house\_value |
| 0.6125498 | 0.1574468 | 1.0000000 | 0.0458573 | 0.0696989 | 0.0517391 | 0.0728618 | 0.1290810 | 1 | 0 | 0 | 0 | 122800 |
| 0.5986056 | 0.1531915 | 0.8823529 | 0.0316966 | 0.0495188 | 0.0184702 | 0.0498355 | 0.1843561 | 1 | 0 | 0 | 0 | 140300 |
| 0.1952191 | 0.5797872 | 0.6274510 | 0.0007647 | 0.0004657 | 0.0005606 | 0.0014803 | 0.2500034 | 0 | 0 | 1 | 0 | 212500 |
| 0.5318725 | 0.3000000 | 0.6666667 | 0.0027688 | 0.0049674 | 0.0132851 | 0.0064145 | 0.0974193 | 0 | 1 | 0 | 0 | 47500 |
| 0.5697211 | 0.1734043 | 0.6274510 | 0.1380993 | 0.1175101 | 0.0571204 | 0.1197368 | 0.4343250 | 1 | 0 | 0 | 0 | 389700 |

## Regressão Linear

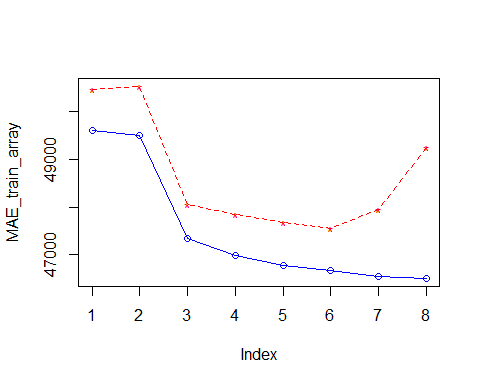
O algoritmo de **Regressão Linear** foi utilizado para predizer os preços dos imóveis onde a variável que se deseja encontrar é a **median\_house\_value** que representa o valor do imóvel baseado em suas features.

As medidas de avaliação utilizadas foram a **MAE (Mean Absolute Error)** que calcula a média da diferença absoluta entre os valores preditos e os observados e o Coeficiente de Determinação (R2) que indica quão bem o modelo consegue se ajustar sobre um conjunto de predições e seus valores verdadeiros, podendo variar entre 0 e 1, sendo que 0 indica que o modelo não consegue explicar a variabilidade dos dados e 1 indica que as predições se ajustam perfeitamente aos dados (modelo explica toda a variabilidade das predições) o que indicará **overfitting**.

## Resultado dos modelos

Avaliação para o modelo **baseline**: MAE **5.047056610^{4}** e R² **63.429319**%.

Avaliação para o modelo **complexo baseado na combinação de features existentes**: MAE **5.09999910^{4}** e R² **63.429319**%.

Avaliação para os modelos **complexos baseado em regressão polinomial**:  
Fórmula grau 1: MAE = **5.047056610^{4}** e R² = **63.429319**%  
Fórmula grau 2: MAE = **5.052782510^{4}** e R² = **63.9071342**%  
Fórmula grau 3: MAE = **4.805290610^{4}** e R² = **66.2598832**%  
Fórmula grau 4: MAE = **4.785344110^{4}** e R² = **66.8109491**%  
Fórmula grau 5: MAE = **4.767135110^{4}** e R² = **66.9997905**%  
Fórmula grau 6: MAE = **4.755851510^{4}** e R² = **67.2089898**%  
Fórmula grau 7: MAE = **4.79587910^{4}** e R² = **61.9227299**%  
Fórmula grau 8: MAE = **4.924575910^{4}** e R² = **-19.8713669**%  
  


A partir da função de grau 7 percebe-se que os modelos tem um bom desempenho nos dados de treinamento, porém não 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 é o polinomial de grau 6, pois na curva do MAE foi o que apresentou menor diferença entre o valor real e o predito durante o treinamento. A escolha da função foi feita de modo a garantir uma convergência entre o treinamento e a validação.

Avaliação para o modelo **polinomial de grau 6**: MAE **4.698425710^{4}** e R² **68.4390569**%.