Desafio Kaggle
Valor de Imóveis

Dados do desafio do Kaggle sobre preço de vendas de imóveis

Link para Kaggle: House Prices: Advanced Regression Techniques

No desafio temos alguns arquivos disponíveis:

- train.csv - o conjunto de treinamento
- test.csv - o conjunto de testes
- data_description.txt - descrição completa de cada coluna
- sample_submission.csv - um envio de benchmark a partir de uma regressão linear no ano e mês da venda, lote quadrado e número de quartos

Base disponível em:
https://raw.githubusercontent.com/Data-Science-FMI/ml-from-scratch-2019/master/data/house_prices_train.csv

A variável "resposta" da base de dados é o Preço da Venda do imóvel "Sale Price"

Passo a passo seguido na análise:

1) Variáveis com *missing data* (valores faltantes)
2) Variáveis com alta concentração de valores
3) Análise de correlação

    Cenário 1: Variáveis Categóricas
    - Entendo o cálculo da Estatística Qui-quadrado: https://math.hws.edu/javamath/ryan/ChiSquare.html
    - Utilizaremos Cramer`s V para medir a associação entre variáveis: https://towardsdatascience.com/the-search-for-categorical-correlation-a1cf7f1888c9 https://www.youtube.com/watch?v=BROyKPwsxKs

    Cenário 2: Variáveis Numéricas
    - Correlação de Pearson para as numéricas

Um ponto para se preocupar é o "vazamento" de uma variável explicativa para a variável resposta:
https://machinelearningmastery.com/data-leakage-machine-learning/

Citar aqui os passos que fizemos para entender se fazia sentido o ajuste da regressão linear.

4) Ajuste de Regressão Linear Simples

Referência teórica
Livro Estatística Básica, WILTON DE O. BUSSAB PEDRO A. MORETTIN
Capítulo 11 - Estimação
Seção: 11.4 Estimadores de Mínimos Quadrados

Referência aplicada
Explicação bem completa com passo a passo didático desde noção de ajuste da reta até análise de resíduos:

https://pt.khanacademy.org/math/statistics-probability/describing-relationships-quantitative-data

Intuição de ajuste de uma reta:

https://pt.khanacademy.org/math/statistics-probability/describing-relationships-quantitative-data/introduction-to-trend-lines/v/fitting-a-line-to-data

- Suposições pro ajuste

A regressão linear simples é uma abordagem linear para modelar o relacionamento entre uma variável explicativa e uma variável resposta, obtendo uma reta que melhor se ajusta aos dados.

$y = a + bx$

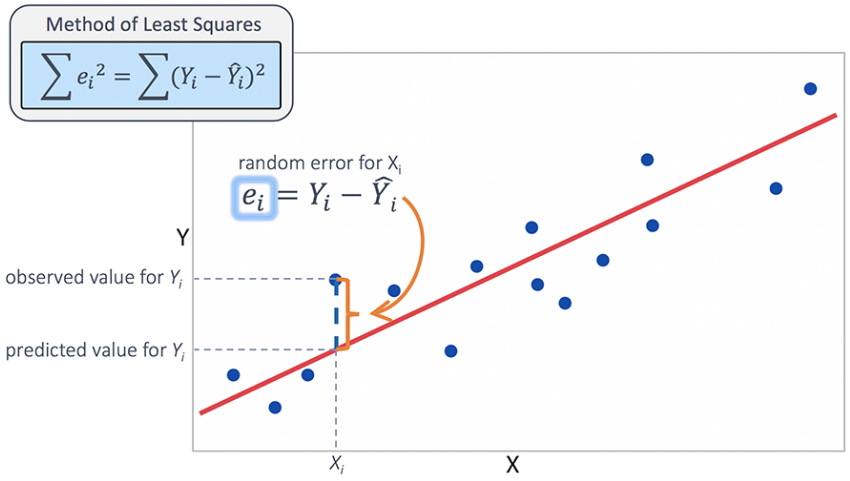
onde:
y é a variável resposta: valor do imóvel,
x é a variável explicativa: que explica o valor do imóvel, por exemplo GrLivArea,

b é a inclinação: se quanto maior o GrLivArea maior for o valor do imóvel, então é algo positivo,
a é a interceptação: que representa o valor do imóvel, quando o GrLivArea é igual a zero.

O objetivo é obter a linha que melhor se ajusta aos nossos dados (a linha que minimiza a soma dos erros quadrados).

O erro é a diferença entre o valor real y e o valor previsto y_hat, que é o valor obtido usando a equação linear calculada.

error = y(real)-y(predito) = y(real)-(a+bx)



- Análise de Resíduos

Aula UFPR sobre análise de resíduos
https://docs.ufpr.br/~niveam/ce071/aula7.pdf

http://www.portalaction.com.br/analise-de-regressao/19-analise-de-residuos-na-regressao-linear-simples

Análise de Variância dos resíduos
https://medium.com/@remycanario17/tests-for-heteroskedasticity-in-python-208a0fdb04ab

https://www.statsmodels.org/stable/generated/statsmodels.stats.diagnostic.het_breuschpagan.html

Análise de Normalidade

Teste de Shapiro-Wilk = scipy.stats.shapiro
https://medium.com/@rrfd/testing-for-normality-applications-with-python-6bf06ed646a9

Teste:
http://www.portalaction.com.br/analise-de-regressao/32-diagnostico-de-homocedasticidade

Gráfico QQPlot: https://pt.wikipedia.org/wiki/Gr%C3%A1fico_Q-Q
https://stackoverflow.com/questions/13865596/quantile-quantile-plot-using-scipy

Tratamento de outliers

- Eliminar o valor: se nosso conjunto de dados é grande o suficiente, poderemos simplesmente deletar os valores anômalos sem maiores prejuízos para a análise.

- Transformação logarítmica: a transformação logarítmica dos dados pode reduzir a variação causada por valores extremos.

- Filtragem de dados: alguns filtros podem ser utilizados, como o média-móvel.

- Tratamento separado: se a quantidade de outliers é significante, podemos tratá-los separadamente. Podemos separar os valores em dois grupos e criar modelos individuais.

Métricas para avaliar o modelo ajustado

https://pt.khanacademy.org/math/ap-statistics/bivariate-data-ap/assessing-fit-least-squares-regression/a/r-squared-intuition

Usando Scikit Learn

Ajuste:
https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

Methods

`fit`(X, y[, sample_weight])	Fit linear model.
`get_params`([deep])	Get parameters for this estimator.
`predict`(X)	Predict using the linear model.
`score`(X, y[, sample_weight])	Return the coefficient of determination R^2 of the prediction.
`set_params`(**params)	Set the parameters of this estimator.

Erro Quadrático Médio:
https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html#sklearn.metrics.mean_squared_error

Fontes:

1)https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.jmp.com%2Fen_ch%2Fstatistics-knowledge-portal%2Fwhat-is-multiple-regression%2Ffitting-multiple-regression-model.html&psig=AOvVaw2B2uAhQdsqJB6T0niJeV7A&ust=1596588041046000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCLD78YyogOsCFQAAAAAdAAAAABAP

2)https://towardsdatascience.com/simple-and-multiple-linear-regression-with-python-c9ab422ec29c

3)https://math.hws.edu/javamath/ryan/ChiSquare.html

Notebook com a implementação passo a passo da regressão linear:
https://colab.research.google.com/drive/1XUpXCGbE9t3_3ECSTGLXDnGavpomVF2D?usp=sharing