 # Machine Learning - Regressão

1. Working Directory

Configurando o diretório de trabalho

setwd("C:/Users/Utilizador/repos/Formacao_cientista_de_dados/big_data_analytics_R_microsoft_azure_machine_lear
getwd()

2. Bussines Problems

Previsão de Despesas Hospitalares

Para esta análise, usaremos um conjunto de dados simulando despesas médicas hipotéticas,

no qual temos um conjunto de pacientes espalhados por 4 regiões do Brasil.

Esse dataset possui 1.338 observações e 7 variáveis.

3. Data Loading despesas <- read.csv("despesas.csv") View(despesas)

4. Descriptive statistics

Descreve, compreende, organiza e resumi os dados

Visualizando as variáveis

str(despesas)

Medias de Tendência Central da variável gastos

summary(despesas$gastos)

Construindo um histograma

hist(despesas$gastos, main = 'Histograma', xlab = 'Gastos')

Tabela de contingência das regiões

table(despesas$regiao)

Explorando relacionamento entre as variáveis: Matriz de Correlação

cor(despesas[c("idade", "bmi", "filhos", "gastos")])

Nenhuma das correlações na matriz é considerada forte, mas existem algumas associações interessantes.

Por exemplo, a idade e o bmi (IMC) parecem ter uma correlação positiva fraca, o que significa que

com o aumento da idade, a massa corporal tende a aumentar.

Há também uma correlação positiva moderada entre a idade e os gastos, além do número de filhos e os gastos.

Estas associações implicam que, à media que idade, massa corporal e número de filhos aumenta, o custo esperado do seguro saúde sobe.

Visualizando relacionamento entre as variáveis: Scatterplot

Perceba que não existe um claro relacionamento entre as variáveis

pairs(despesas[c("idade", "bmi", "filhos", "gastos")])

5. Scatterplot Matrix install.packages("mnormt") install.packages("psych") library(psych)

Este gráfico fornece mais informações sobre o relacionamento entre as variáveis

pairs.panels(despesas[c("idade", "bmi", "filhos", "gastos")])

6. Training the Model (using training data)

Treinando o Modelo (usando os dados de treino)

variavel target (dependente, quero prever): gastos. Lado esquerdo do ~ é a variavel target

variaveis preditoras: idade, filhos, bmi, sexo, fumante e região. Lado direito do ~

Regressão Linear Multipla: várias variavéis preditoras

Regressão Linear Simple: uma variavél preditora

modelo <- lm(gastos ~ idade + filhos + bmi + sexo + fumante + regiao, data = despesas)

Similar ao item anterior: outra forma de fazer

modelo <- lm(gastos ~ ., data = despesas)

Visualizando os coeficientes

modelo

6.1 Prediction

Prevendo despesas médicas

?predict

Aqui verificamos os gastos previstos pelo modelo que devem ser iguais aos dados de treino

previsao1 <- predict(modelo) View(previsao1)

6.2 Training data forecast

Prevendo os gastos com Dados de teste: data set de treino

despesasteste <- read.csv("despesas-teste.csv")

ver dataset

View(despesasteste)

Previsão 2 com dados de teste

previsao2 <- predict(modelo, despesasteste)

ver previsão

View(previsao2)

7. Evaluating the Model's Performance

Etapa 4: Avaliando a Performance do Modelo

Mais detalhes sobre o modelo

summary(modelo)

****

***** Estas informações abaixo é que farão de você *****

***** um verdadeiro conhecedor de Machine Learning *****

****

Equação de Regressão

y = a + bx (simples)

y = a + b0x0 + b1x1 (múltipla)

Resíduos (Residuals)

Diferença entre os valores observados de uma variável e seus valores previstos

Seus resíduos devem se parecer com uma distribuição normal, o que indica

que a média entre os valores previstos e os valores observados é próximo de 0 (o que é bom)

Coeficiente - Intercept - a (alfa)

Valor de a na equação de regressão

Coeficientes - Nomes das variáveis - b (beta)

Valor de b na equação de regressão

Obs: A questão é que lm() ou summary() têm diferentes convenções de

rotulagem para cada variável explicativa.

Em vez de escrever slope_1, slope_2, ...

Eles simplesmente usam o nome da variável em qualquer saída para

indicar quais coeficientes pertencem a qual variável.

Etapa 5: Otimizando a Performance do Modelo

Adicionando uma variável com o dobro do valor das idades

despesas$idade2 <- despesas$idade ^ 2$

Adicionando um indicador para BMI >= 30

despesas$bmi30 <- ifelse(despesas$bmi >= 30, 1, 0)$

View(despesas)

Criando o modelo final

modelo_v2 <- lm(gastos ~ idade + idade2 + filhos + bmi + sexo + bmi30 * fumante + regiao, data = despesas)

summary(modelo_v2)

Dados de teste

despesasteste <- read.csv("despesas-teste.csv") View(despesasteste)

previsao <- predict(modelo, despesasteste) class(previsao) View(previsao)