**PrevendoDespesasHospitalares.r**

Rafael

Sun Mar 17 16:19:42 2019

#Machine Learning - Regressão - Prevendo Despesas Hospitalares   
  
*# Configurando o diretório de trabalho*  
*# Coloque entre aspas o diretório de trabalho que você está usando no seu computador*  
*#setwd("{SET\_YOUR\_HOME\_DIRECTORY\_HERE}")*  
**getwd**()

## [1] "C:/WorkSpace/MachineLearning\_Linear\_Regression\_R"

*# Problema de Negócio: Previsão de Despesas Hospitalares*  
  
*# Para esta análise, vamos usar um conjunto de dados simulando despesas médicas hipotéticas*   
*# para um conjunto de pacientes espalhados por 4 regiões do Brasil.*  
*# Esse dataset possui 1.338 observações e 7 variáveis.*  
  
*# Observação Importante: Nesse estudo de caso não estamos separando os dados em Dataset de Treino e Teste.*  
  
##################################################################### *# Etapa 1 - Coletando os dados*  
#####################################################################  
despesas <- **read.csv**("despesas.csv")  
**View**(despesas)  
  
##################################################################### *# Etapa 2: Explorando e Preparando os Dados*  
#####################################################################  
*# Visualizando as variáveis*  
**str**(despesas)

## 'data.frame': 1338 obs. of 7 variables:  
## $ idade : int 19 18 28 33 32 31 46 37 37 60 ...  
## $ sexo : Factor w/ 2 levels "homem","mulher": 2 1 1 1 1 2 2 2 1 2 ...  
## $ bmi : num 27.9 33.8 33 22.7 28.9 25.7 33.4 27.7 29.8 25.8 ...  
## $ filhos : int 0 1 3 0 0 0 1 3 2 0 ...  
## $ fumante: Factor w/ 2 levels "nao","sim": 2 1 1 1 1 1 1 1 1 1 ...  
## $ regiao : Factor w/ 4 levels "nordeste","norte",..: 3 4 4 1 1 4 4 1 2 1 ...  
## $ gastos : num 16885 1726 4449 21984 3867 ...

*# Medias de Tendência Central da variável gastos*  
**summary**(despesas**$**gastos)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1122 4740 9382 13270 16640 63770

*# Construindo um histograma*  
**hist**(despesas**$**gastos,   
 main = 'Histograma',   
 xlab = 'Gastos')



*# Tabela de contingência das regiões*  
**table**(despesas**$**regiao)

##   
## nordeste norte sudeste sul   
## 325 324 325 364

*# Explorando relacionamento entre as variáveis: Matriz de Correlação*  
**cor**(despesas[**c**("idade", "bmi", "filhos", "gastos")])

## idade bmi filhos gastos  
## idade 1.0000000 0.10934101 0.04246900 0.29900819  
## bmi 0.1093410 1.00000000 0.01264471 0.19857626  
## filhos 0.0424690 0.01264471 1.00000000 0.06799823  
## gastos 0.2990082 0.19857626 0.06799823 1.00000000

*# 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")])  
  
*# Scatterplot Matrix {Pacote bastente util para visualizar uma mátrix de gráficos}*  
*#install.packages("psych")*  
**library**(psych)

## Warning: package 'psych' was built under R version 3.5.3



*# Este gráfico fornece mais informações sobre o relacionamento entre as variáveis*  
**pairs.panels**(despesas[**c**("idade", "bmi", "filhos", "gastos")])



##################################################################### *# Etapa 3: Treinando o Modelo (No caso de uso não estamos separando em dados de treino e teste)*  
#####################################################################  
*# Lado Esquerdo do ~ -> Target (variavel dependente: Y)*  
*# Lado Direito do ~ -> Variáveis preditores (Variavel independente (X))*  
*# Data -> DataSet a ser analisado*  
*# Regressão Linear Multiplas (+ de 1 Variavel Preditora)*  
modelo <- **lm**(gastos **~** ., data = despesas) *# . representa todas variaveis*  
  
*# Visualizando os coeficientes*  
modelo

##   
## Call:  
## lm(formula = gastos ~ ., data = despesas)  
##   
## Coefficients:  
## (Intercept) idade sexomulher bmi filhos   
## -12425.7 256.8 131.4 339.3 475.7   
## fumantesim regiaonorte regiaosudeste regiaosul   
## 23847.5 352.8 -606.5 -682.8

*# Prevendo despesas médicas utilizando modelos lineares*  
*# Aqui verificamos os gastos previstos pelo modelo que devem ser iguais aos dados de treino*  
previsao1 <- **predict**(modelo)  
**View**(previsao1)  
  
*# Prevendo os gastos com Dados de teste*  
despesasteste <- **read.csv**("despesas-teste.csv") *#Dados que o modelo não conhece*  
**View**(despesasteste)  
  
previsao2 <- **predict**(modelo, despesasteste)  
**View**(previsao2)  
  
  
##################################################################### *# Etapa 4: Avaliando a Performance do Modelo*  
#####################################################################  
*# Mais detalhes sobre o modelo*  
**summary**(modelo)

##   
## Call:  
## lm(formula = gastos ~ ., data = despesas)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11302.7 -2850.9 -979.6 1383.9 29981.7   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -12425.7 1000.7 -12.418 < 2e-16 \*\*\*  
## idade 256.8 11.9 21.586 < 2e-16 \*\*\*  
## sexomulher 131.3 332.9 0.395 0.693255   
## bmi 339.3 28.6 11.864 < 2e-16 \*\*\*  
## filhos 475.7 137.8 3.452 0.000574 \*\*\*  
## fumantesim 23847.5 413.1 57.723 < 2e-16 \*\*\*  
## regiaonorte 352.8 476.3 0.741 0.458976   
## regiaosudeste -606.5 477.2 -1.271 0.203940   
## regiaosul -682.8 478.9 -1.426 0.154211   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6062 on 1329 degrees of freedom  
## Multiple R-squared: 0.7509, Adjusted R-squared: 0.7494   
## F-statistic: 500.9 on 8 and 1329 DF, p-value: < 2.2e-16

*# \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\**  
*# \*\*\* Informações abaixo são de suma importância para compreender o modelo treinado \*\*\**  
*# \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\**  
  
*# Equação de Regressão*  
*# y = a + bx (simples)*  
*# y = a + b0x0 + b1x1 (múltipla)*  
  
*# Resíduos*  
*# 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) Representa onde a reta de regressão corta o eixo Y.*  
*# Valor de a na equação de regressão*  
  
*# Coeficientes - Nomes das variáveis - b (beta) Slope Inclinação da Reta*  
*# 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.*  
  
*# Erro Padrão Std. Error*  
*# Medida de variabilidade na estimativa do coeficiente a (alfa). O ideal é que este valor*   
*# seja menor que o valor do coeficiente, mas nem sempre isso irá ocorrer.*  
  
*# Asteriscos*   
*# Os asteriscos representam os níveis de significância de acordo com o p-value.*  
*# Quanto mais estrelas, maior a significância.*  
*# Atenção --> Muitos astericos indicam que é improvável que não exista*   
*# relacionamento entre as variáveis.*  
  
*# Valor t*  
*# Define se coeficiente da variável é significativo ou não para o modelo.*   
*# Ele é usado para calcular o p-value e os níveis de significância.*  
  
*# p-value (Pr(>|t|))*  
*# O p-value representa a probabilidade que a variável não seja relevante.*   
*# Deve ser o menor valor possível.*   
*# Se este valor for realmente pequeno, o R irá mostrar o valor*   
*# como notação científica*  
  
*# Significância*  
*# São aquelas legendas próximas as suas variáveis (Legenda dos asteriscos)*  
*# Espaço em branco - ruim (Sem Relevância, pode-se remover do modelo)*  
*# Pontos - razoável*   
*# Asteriscos - bom*  
*# Muitos asteriscos - muito bom*  
  
*# Residual Standar Error*  
*# Este valor representa o desvio padrão dos resíduos*  
  
*# Degrees of Freedom*  
*# É a diferença entre o número de observações na amostra de treinamento*   
*# e o número de variáveis no seu modelo*  
  
*# R-squared (coeficiente de determinação - R^2)*  
*# Ajuda a avaliar o nível de precisão do nosso modelo.*   
*# Quanto maior, melhor, sendo 1 o valor ideal.*  
  
*# F-statistics*  
*# É o teste F do modelo. Esse teste obtém os parâmetros do nosso modelo*   
*# e compara com um modelo que tenha menos parâmetros.*  
*# Em teoria, um modelo com mais parâmetros tem um desempenho melhor.*   
  
*# Se o seu modelo com mais parâmetros NÃO tiver perfomance*  
*# melhor que um modelo com menos parâmetros, o valor do p-value será bem alto.*   
  
*# Se o modelo com mais parâmetros tiver performance*  
*# melhor que um modelo com menos parâmetros, o valor do p-value será mais baixo.*  
  
*# Lembre-se que correlação não implica causalidade*  
  
  
##################################################################### *#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*  
*#?ifelse*  
despesas**$**bmi30 <- **ifelse**(despesas**$**bmi **>=** 30, 1, 0) *#Se for >=30 coloca 1, senão 0*  
  
**View**(despesas)  
  
*# Criando o modelo final*  
modelo\_v2 <- **lm**(gastos **~** idade **+** idade2 **+** filhos **+** bmi **+** sexo **+**  
 (bmi30 **\*** fumante) **+** regiao, data = despesas)  
  
  
*# É possível verificar que com as alterações dos inputs o modelo\_v2 obteve uma melhor taxa de acerto (87%)*  
*# em comparação ao modelo (75%)*  
**summary**(modelo\_v2)

##   
## Call:  
## lm(formula = gastos ~ idade + idade2 + filhos + bmi + sexo +   
## (bmi30 \* fumante) + regiao, data = despesas)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -17297.1 -1656.0 -1262.7 -727.8 24161.6   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -636.9298 1361.0589 -0.468 0.639886   
## idade -32.6181 59.8250 -0.545 0.585690   
## idade2 3.7307 0.7463 4.999 6.54e-07 \*\*\*  
## filhos 678.6017 105.8855 6.409 2.03e-10 \*\*\*  
## bmi 119.7715 34.2796 3.494 0.000492 \*\*\*  
## sexomulher 496.7690 244.3713 2.033 0.042267 \*   
## bmi30 -997.9355 422.9607 -2.359 0.018449 \*   
## fumantesim 13404.5952 439.9591 30.468 < 2e-16 \*\*\*  
## regiaonorte 279.1661 349.2826 0.799 0.424285   
## regiaosudeste -942.9958 350.1754 -2.693 0.007172 \*\*   
## regiaosul -548.8684 352.1950 -1.558 0.119372   
## bmi30:fumantesim 19810.1534 604.6769 32.762 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4445 on 1326 degrees of freedom  
## Multiple R-squared: 0.8664, Adjusted R-squared: 0.8653   
## F-statistic: 781.7 on 11 and 1326 DF, p-value: < 2.2e-16

**summary**(modelo)

##   
## Call:  
## lm(formula = gastos ~ ., data = despesas)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11302.7 -2850.9 -979.6 1383.9 29981.7   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -12425.7 1000.7 -12.418 < 2e-16 \*\*\*  
## idade 256.8 11.9 21.586 < 2e-16 \*\*\*  
## sexomulher 131.3 332.9 0.395 0.693255   
## bmi 339.3 28.6 11.864 < 2e-16 \*\*\*  
## filhos 475.7 137.8 3.452 0.000574 \*\*\*  
## fumantesim 23847.5 413.1 57.723 < 2e-16 \*\*\*  
## regiaonorte 352.8 476.3 0.741 0.458976   
## regiaosudeste -606.5 477.2 -1.271 0.203940   
## regiaosul -682.8 478.9 -1.426 0.154211   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6062 on 1329 degrees of freedom  
## Multiple R-squared: 0.7509, Adjusted R-squared: 0.7494   
## F-statistic: 500.9 on 8 and 1329 DF, p-value: < 2.2e-16

*# Os ajustes realizados no dataset "despesas" para treinar o modelo, não foram realizados no dataset "despesas-teste"*  
*# Para executar a previsão de valores nos dados do dataset "despesas-teste" deve-se incluir as novas colunas no dataset*  
*# igual foi feito ao dataset "despesas".*  
despesasteste <- **read.csv**("despesas-teste.csv")  
despesasteste**$**idade2 <- despesasteste**$**idade **^** 2  
  
despesasteste**$**bmi30 <- **ifelse**(despesasteste**$**bmi **>=** 30, 1, 0) *#Se for >=30 coloca 1, senão 0*  
  
previsaofinal <- **predict**(modelo\_v2, despesasteste)  
**View**(previsaofinal)  
  
*# Combinando o dataset "despesasteste" com a previsão resolvida pelo modelo linear*  
DataSetFinal <- **cbind**(despesasteste,previsaofinal)  
**View**(DataSetFinal)