Tech Challenge I

IA para Devs

Daniela Cruz de Malta - 353365

Gabriela Maciel Godoi - 355125

Lucas Sutelo - 353721

28 de Maio de 2024

Sumário

Caso de estudo

Nesse primeiro Tech Challenge o objetivo foi aplicar um modelo de regressão múltipla para prever custos com o plano de saúde. Esse é um ótimo exemplo de estudo, pois diversas variáveis podem contribuir com esse custo, como idade, IMC, número de filhos, e etc.

Exploração de dados

1. **A base de dados**

O grupo decidiu utilizar a base de dados fornecida no Tech Challenge, sendo que complementamos a base utilizando o código que está no anexo.

O resultado foi uma base de dados com 21336 linhas e 7 colunas contendo idade, gênero, *BMI* (que é o IMC), quantidade de filhos, se é fumante, região e os custos.

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520
...	...	...	...	...	...	...	...
285	46	male	26.620	1	no	southeast	7742.10980
286	46	female	48.070	2	no	northeast	9432.92530
287	63	female	26.220	0	no	northwest	14256.19280
288	59	female	36.765	1	yes	northeast	47896.79135
289	52	male	26.400	3	no	southeast	25992.82104

Imagem 1: Foto da base de dados

A próxima imagem mostra o perfil desses dados para as variáveis numéricas. Podemos ver, por exemplo, que o custo mínimo é de 1121 dólares e o máximo é 63770 dólares, o que nos ajuda a ter uma dimensão do valor do erro aceito para o nosso modelo. No código também verificamos que não existem dados nulos na base de dados.

	age	bmi	children	charges
count	21336.000000	21336.000000	21336.000000	21336.000000
mean	39.242501	30.647658	1.096269	13284.370438
std	14.032129	6.069989	1.205142	12101.379287
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.272500	0.000000	4747.052900
50%	39.000000	30.400000	1.000000	9386.161300
75%	51.000000	34.700000	2.000000	16657.717450
max	64.000000	52.580000	5.000000	63770.428010

Imagem 2: Perfil da base de dados.

2. Explorando a base de dados

Na análise exploratória o grupo fez uso principalmente do gráfico *boxplot* que mostra a simetria dos dados, a mediana, os limites superiores e inferiores e os dados que parecem ser *outliers*. Além disso foram utilizados alguns gráficos de histograma que mostram como está a distribuição dos dados.

A variável idade (ou Age) não possui *outliers* e é uma variável bem simétrica, sendo que existe uma grande concentração nos dados abaixo dos 20 anos.

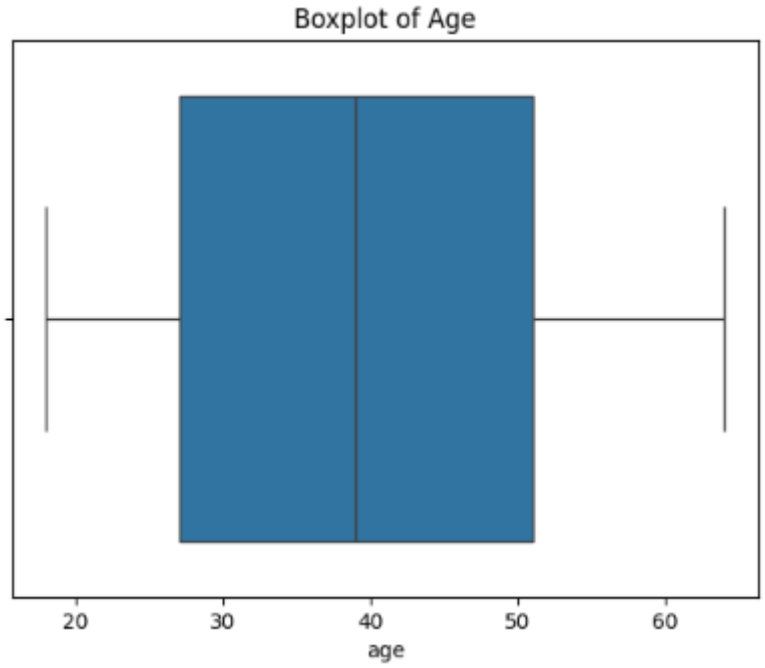

Imagem 3: *Boxplot* da variável idade

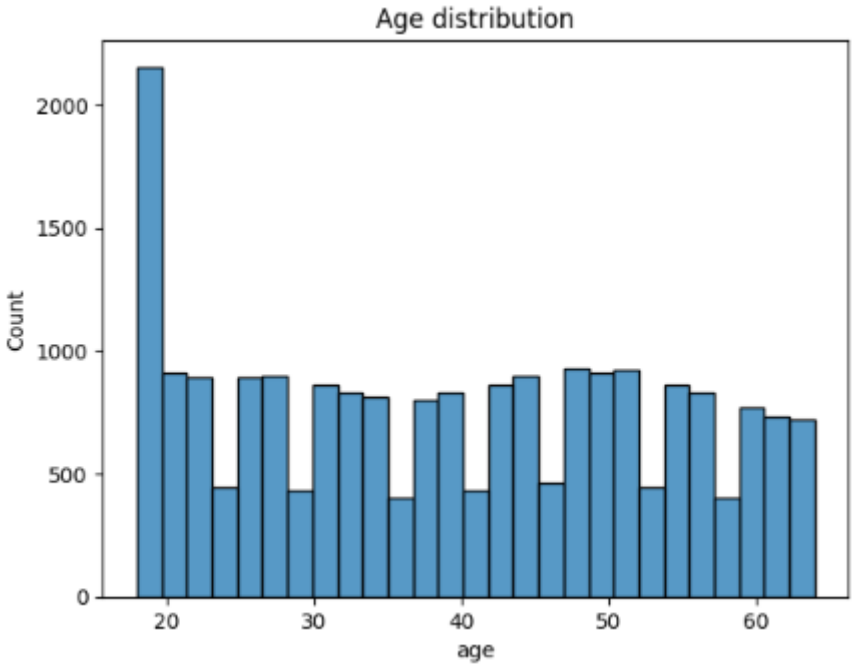

Imagem 4: Distribuição da variável idade

O IMC (ou *bmi*) é uma variável menos simétrica, se verificarmos nas tabelas de IMC vemos que a maior parte das pessoas nessa base de dados está acima do peso.

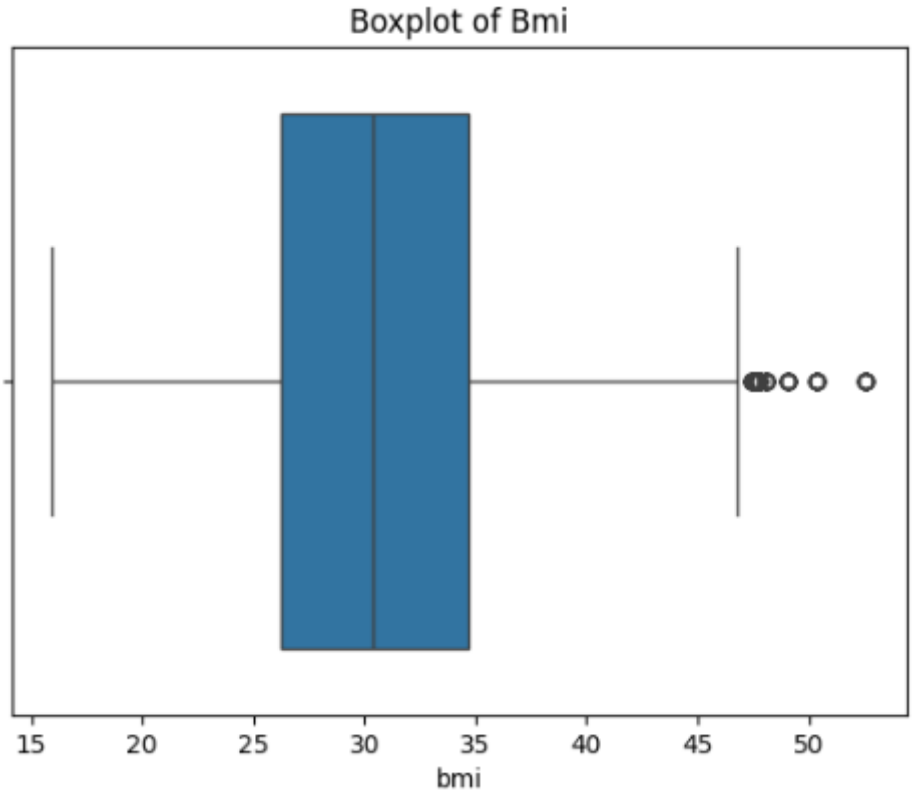

Imagem 5: *Boxplot* da variável IMC

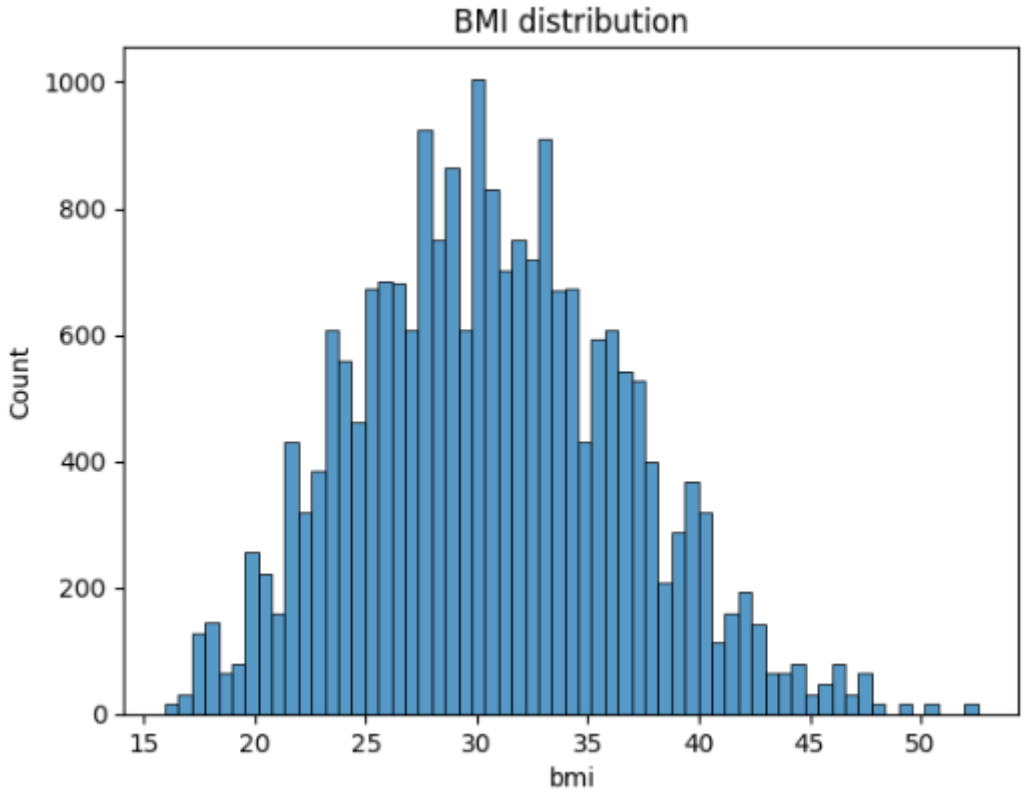

Imagem 6: Distribuição da variável IMC

A quantidade de não fumantes é maior que a quantidade de fumantes nessa base de dados, sendo que isso não varia com o gênero.

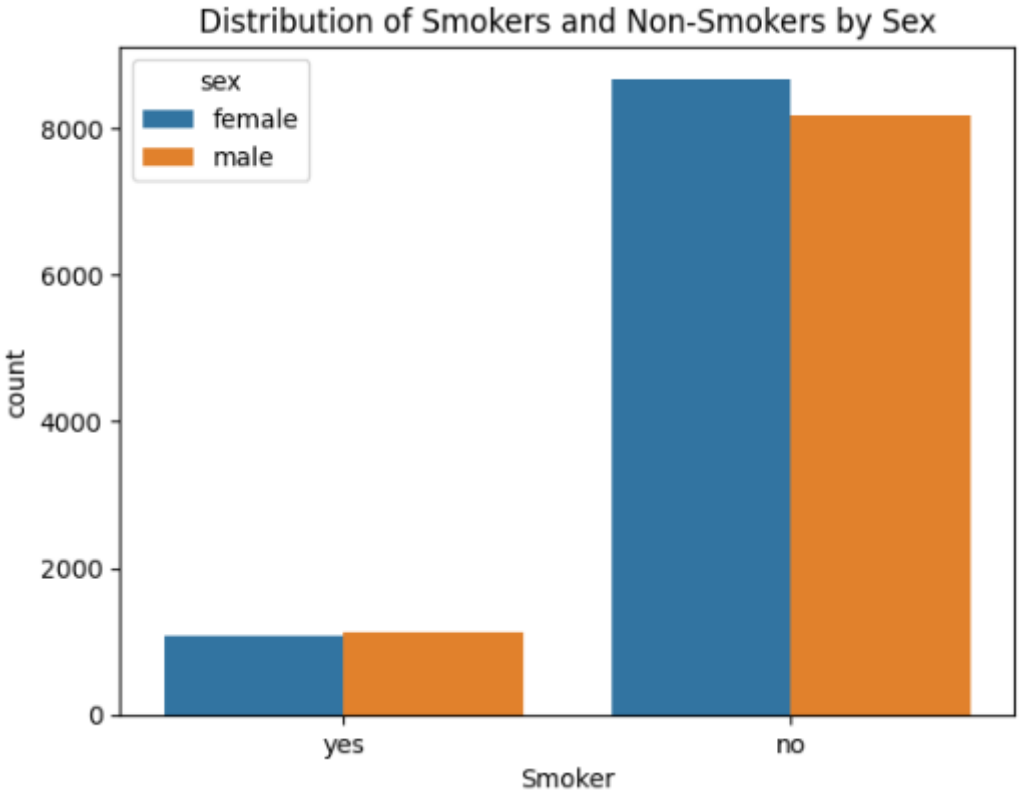

Imagem 7: Quantidade de fumantes

Verificando a variável de custos, vemos que alguns dos custos estão bem acima da média, mas é algo que pode ser explicado pelas demais variáveis.

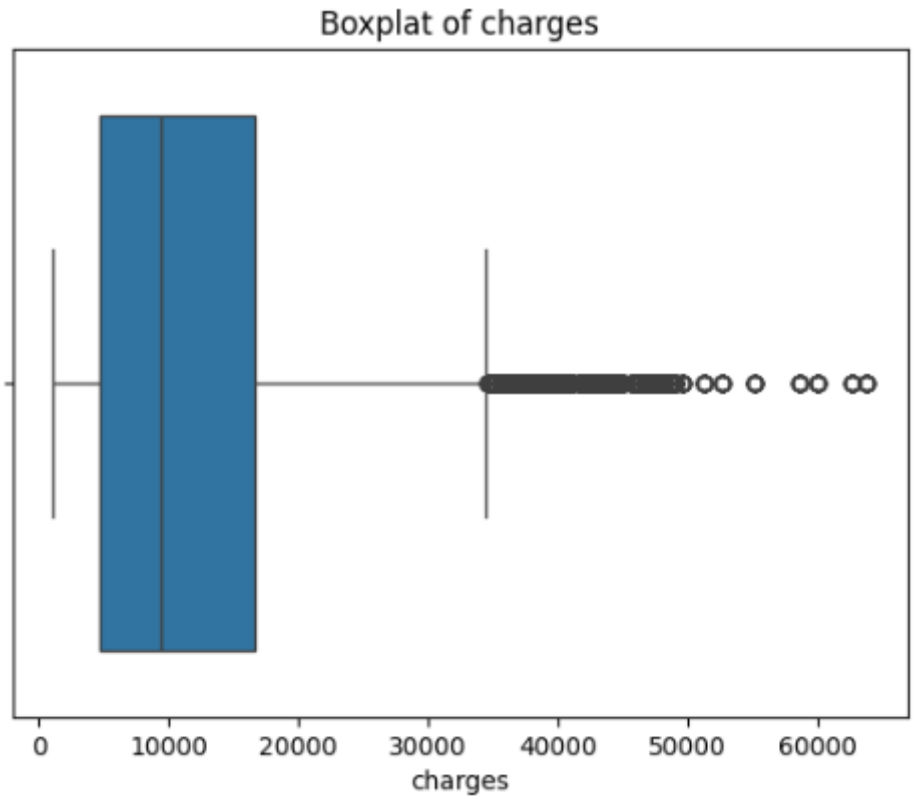

Imagem 8: *Boxplot* da variável custo

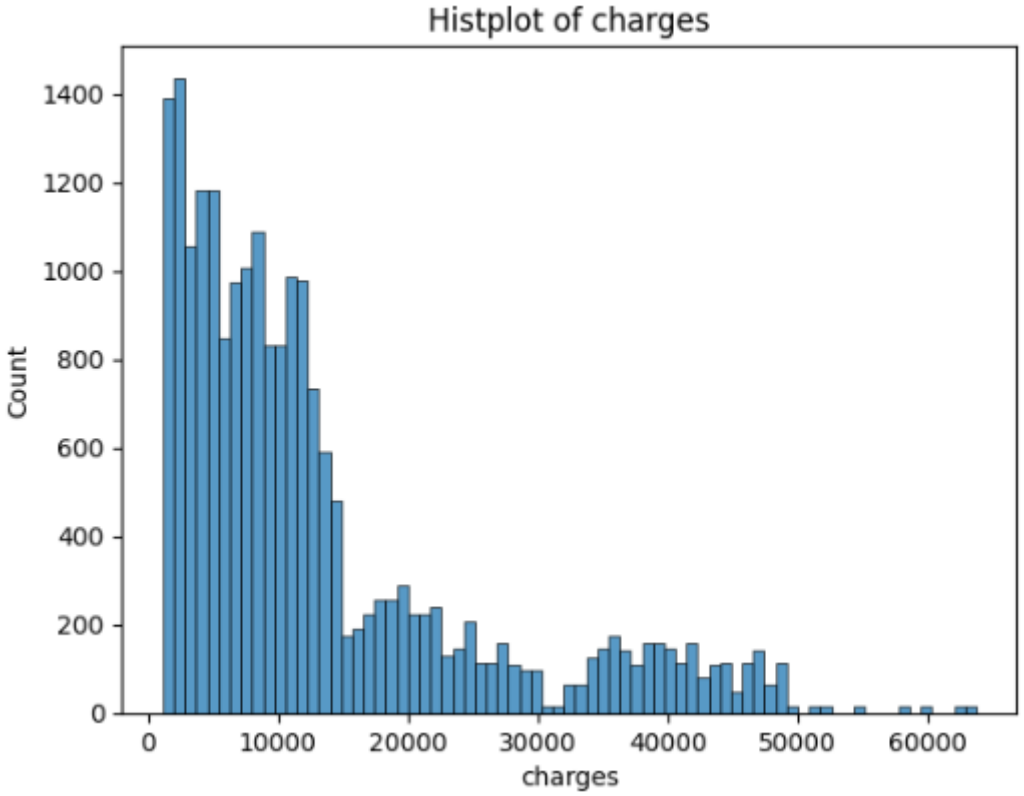

Imagem 9: Distribuição da variável custo

Pré-processamento de dados

O tratamento de dados realizado nesta base foi a aplicação do Label Encoder para converter as variáveis não numéricas e numéricas e a normalização para melhorar os resultados do modelo.

	age	sex	bmi	children	smoker	region	charges
0	19	0	27.900	0	1	3	16884.92400
1	18	1	33.770	1	0	2	1725.55230
2	28	1	33.000	3	0	2	4449.46200
3	33	1	22.705	0	0	1	21984.47061
4	32	1	28.880	0	0	1	3866.85520

Imagem 10: Base de dados após a aplicação do Label Encoder

Com todos os dados numéricos foi possível realizar a correlação entre as variáveis. Veja que a variável mais relacionada com o custo é a de fumante, o que pode nos mostrar que a pessoa ser fumante pode ter uma relação com custos maiores no plano de saúde. Outras variáveis que se correlacionam com o custo são o IMC e a idade, mas com menos força que o tabagismo.

Imagem 11: Correlação entre as variáveis

Agora sim, podemos separar as bases de treino e teste para partir para aplicação do modelo de regressão múltipla. A proporção de 20% para dados de teste foi usada, sendo que a base de treino ficou com 14295 linhas e a base de teste ficou com 7041 linhas.

Somente após a separação das bases podemos aplicar a normalização (de modo que os dados não fiquem enviesados). Os resultados dessa normalização podem ser vistos no exemplo com os dados de teste nas imagens a seguir.

	age	sex	bmi	children	smoker	region
14613	34	0	27.720	0	0	2
1370	28	1	36.400	1	1	3
3090	64	1	39.160	1	0	2
18338	34	0	23.560	0	0	0
5038	55	0	35.200	0	1	2
...	...	...	...	...	...	...
11284	34	0	33.250	1	0	0
11964	32	0	29.735	0	0	1
5390	18	0	38.665	2	0	0
860	38	0	28.000	3	0	3
15795	33	0	19.095	2	1	0

Imagem 12: Dados de teste antes da normalização

	age	sex	bmi	children	smoker	region
0	0.347826	0.0	0.321136	0.0	0.0	0.666667
1	0.217391	1.0	0.558165	0.2	1.0	1.000000
2	1.000000	1.0	0.633534	0.2	0.0	0.666667
3	0.347826	0.0	0.207537	0.0	0.0	0.000000
4	0.804348	0.0	0.525396	0.0	1.0	0.666667
...	...	...	...	...	...	...
14290	0.347826	0.0	0.472146	0.2	0.0	0.000000
14291	0.304348	0.0	0.376161	0.0	0.0	0.333333
14292	0.000000	0.0	0.620016	0.4	0.0	0.000000
14293	0.434783	0.0	0.328782	0.6	0.0	1.000000
14294	0.326087	0.0	0.085609	0.4	1.0	0.000000

Imagem 13: Dados de teste após a normalização

Modelo de regressão múltipla

Aplicamos o modelo de regressão múltipla aos dados tratados. O modelo de regressão simples utiliza a função de primeiro grau (y = ax + b) para encontrar a melhor reta para a base de dados, ou seja, a reta que gera o menor erro versus os dados reais. No modelo de regressão múltipla se tenta encontrar a equação de reta, porém com mais coeficientes a_n e b_n, cada um para uma variável da base de dados.

Veja a seguir os resultados obtidos e conclusões sobre esses resultados.

```
P-values do modelo:
const       0.000000e+00
age         0.000000e+00
sex         1.753111e-01
bmi         0.000000e+00
children    1.848307e-29
smoker      0.000000e+00
region      7.680618e-12
dtype: float64
```

Imagem 14: Resultado dos P-values.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                charges   R-squared:                       0.751
Model:                            OLS   Adj. R-squared:                  0.751
Method:                 Least Squares   F-statistic:                     7177.
Date:                Tue, 28 May 2024   Prob (F-statistic):               0.00
Time:                        00:29:17   Log-Likelihood:            -1.4476e+05
No. Observations:               14295   AIC:                         2.895e+05
Df Residuals:                   14288   BIC:                         2.896e+05
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -1.191e+04    291.207    -40.884      0.000   -1.25e+04   -1.13e+04
age          258.0688      3.630     71.088      0.000     250.953     265.185
sex         -137.8145    101.678     -1.355      0.175    -337.117      61.488
bmi          334.0621      8.515     39.233      0.000     317.372     350.752
children     475.0368     42.057     11.295      0.000     392.600     557.473
smoker      2.381e+04    125.717    189.369      0.000    2.36e+04    2.41e+04
region      -317.7522     46.386     -6.850      0.000    -408.676    -226.829
==============================================================================
Omnibus:                     3104.393   Durbin-Watson:                   2.001
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             7549.071
Skew:                           1.205   Prob(JB):                         0.00
Kurtosis:                       5.621   Cond. No.                         296.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Imagem 15: Resultado sumarizado

	2.5%	97.5%
const	-12476.628201	-11335.020819
age	250.953003	265.184627
sex	-337.116528	61.487584
bmi	317.372015	350.752275
children	392.600268	557.473247
smoker	23560.586055	24053.430867
region	-408.675606	-226.828870

Imagem 16: Intervalos de confiança do modelo

```
MAE: 4169.067480392558
MSE: 36313184.985052556
RMSE: 6026.042232265931
```

Imagem 17: Erros do modelo

A principal métrica que podemos usar para avaliar a performance do nosso modelo é o R-score que ficou em 0.751, o que pode ser um bom resultado tendo em vista a quantidade de dados de treino e que não foi feito nenhum tratamento adicional nesses dados. Uma possível melhoria futura seria complementar e melhorar a base de dados para que o resultado seja acima de 0.8.

Nos intervalos de confiança podemos verificar o quanto o custo aumenta ou diminui de acordo com as demais variáveis. Veja, por exemplo, que a variável fumante aumenta muito o custo (intervalo mínimo de 23560 dólares), já a variável idade tem uma influência menor (intervalo máximo de 265 dólares).

Nos erros podemos verificar que o modelo tem um erro médio de 4169 dólares e que a raiz do erro quadrático é 6026 dólares. Uma melhoria para esse modelo seria diminuir a diferença entre esses erros, pois uma menor diferença significa ter uma menor influência dos *outliers* da base de teste.

Conclusão

Com este Tech Challenge o grupo conseguiu colocar em prática os assuntos abordados durante a primeira fase da pós, que foi bem focada em modelos de Machine Learning, que é a base para os modelos de inteligência artificial.

O modelo de regressão é amplamente utilizado em diversos setores e casos. Pode ser usado, por exemplo, para fazer previsões de vendas, encontrar possíveis valores de imóveis, entre outros.

Verificamos os passos necessários para criar um modelo e, principalmente, como avaliar os resultados do modelo para que possa ser usado para fazer novas previsões.

Anexo

Código utilizado para complementar a base de dados.

```
import pandas as pd
import numpy as np

# Carregar o conjunto de dados recém-carregado
new_file_path = "custos_medicos.csv"
new_df = pd.read_csv(new_file_path)

# Gerar dados adicionais
num_new_rows = 10000
np.random.seed(42)

# Gerar novos dados
add_data = {
    'age': np.random.randint(18, 80, num_new_rows),
    'sex': np.random.choice(['male', 'female'], num_new_rows),
    'bmi': np.random.uniform(15, 40, num_new_rows),
    'children': np.random.randint(0, 5, num_new_rows),
    'smoker': np.random.choice(['yes', 'no'], num_new_rows),
    'region': np.random.choice(['north', 'south', 'east', 'west', 'southwest'], num_new_rows),
    'charges': np.random.uniform(1000, 50000, num_new_rows)
}

# Criar DataFrame com os novos dados
add_df = pd.DataFrame(add_data)

# Anexar os novos dados ao conjunto de dados existente
combined_df = pd.concat([new_df, add_data], ignore_index=True)

# Salvar o conjunto de dados atualizado em CSV
combined_file = 'path_to_your_file/medical_costs_updated_with_additional.csv'
combined_df.to_csv(combined_file, index=False)
```