Tech Challenge – Fase 1

A proposta do desafio era desenvolver um modelo preditivo de regressão para prever o valor dos custos médicos individuais cobrados pelo seguro de saúde.

A base de dados escolhida foi encontrada no Kaggle e está disponível no github juntamente com o notebook desenvolvido.

<https://github.com/jubasoler/tech_challenge-fase1>

<https://www.google.com/url?q=https%3A%2F%2Fwww.kaggle.com%2Fdatasets%2Fmragpavank%2Finsurance1>

**Exploração de dados**

Inicialmente realizei a importação da base de dados utilizando a biblioteca pandas e passei a fazer a análise exploratória dos dados.

Na análise dos dados pude observar que a base possui 1338 linhas, 6 features, nenhuma possui valores nulos e 3 delas (sex, smoker e region) são do tipo texto. Nossa variável Target é a classe **charges** (encargos/cobranças).

**Features**

* Idade
* Gênero
* IMC
* Quantidade de Filhos
* Fumante
* Região

Como três features são categóricas, utilizei a função ***value\_counts*** para verificar quantas categorias existiam em cada uma.

Gráfico, Gráfico de pizza

Descrição gerada automaticamente Gráfico, Gráfico de pizza

Descrição gerada automaticamente Gráfico, Gráfico de pizza

Descrição gerada automaticamente

Pude observar que os pacientes da amostra estão aproximadamente distribuídos uniformemente entre as 4 regiões, a maioria são não fumantes e tem praticamente a mesma quantidade de homens e mulheres.

Em seguida utilizei a função ***describe*** nas features numéricas, que me trouxe os seguintes resultados:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **age** | **bmi** | **children** | **charges** |
| **count** | 1338.000000 | 1338.000000 | 1338.000000 | 1338.000000 |
| **mean** | 39.207025 | 30.663397 | 1.094918 | 13270.422265 |
| **std** | 14.049960 | 6.098187 | 1.205493 | 12110.011237 |
| **min** | 18.000000 | 15.960000 | 0.000000 | 1121.873900 |
| **25%** | 27.000000 | 26.296250 | 0.000000 | 4740.287150 |
| **50%** | 39.000000 | 30.400000 | 1.000000 | 9382.033000 |
| **75%** | 51.000000 | 34.693750 | 2.000000 | 16639.912515 |
| **max** | 64.000000 | 53.130000 | 5.000000 | 63770.428010 |

Percebe-se que:

* A média de idade dos pacientes é de 39 anos, variando de 18 a 64 anos;
* O IMC varia de 15,96 a 53,13 com média de 30,66;
* Existem pacientes com nenhum filho até paciente com 5 filhos;
* Os custos hospitalares têm uma média de 13270.

Histogramas

Gráfico, Histograma

Descrição gerada automaticamente

Gráfico, Gráfico de barras, Gráfico de caixa estreita

Descrição gerada automaticamente Gráfico

Descrição gerada automaticamente

Gráfico, Gráfico de barras

Descrição gerada automaticamente

Observando especialmente o diagrama de caixa dos custos hospitalares percebe-se a presença de outliers.

**Pré-processamento de dados**

Feita a análise exploratória iniciei a preparação dos dados para o modelo e verifiquei que as escalas dos dados eram muito diferentes e resolvi aplicar as seguintes técnicas de transformação nos dados:

* Label Enconder para tratar variáveis categóricas que possuem apenas dois tipos de dados como sexo e se é fumante;
* One Hot Encoding para a região.

Após o pré-processamento gerei a matriz de correlação e pude observar que as features mais promissoras para prever o custo hospitalar eram as classes **smoker** e **bmi**.

Gráfico

Descrição gerada automaticamente

Plotei, então, essas features em um gráfico de scatter para analisar com mais detalhes:

Gráfico, Gráfico de dispersão

Descrição gerada automaticamente

A correlação realmente mostra uma certa tendência ascendente nos dados.

**Modelagem e Validação**

Na fase de modelagem utilizei 3 técnicas para verificar qual se encaixava melhor:

* Linear Regression
* Decision Tree Regression
* Random Forest Regression

Para tanto, dividi as bases em conjuntos de treinamento e teste, deixando 20% para os testes e fiz a validação com dados normais e dados padronizados com **StandardScaler**. Abaixo apresento os resultados obtidos em cada modelo.

Interface gráfica do usuário

Descrição gerada automaticamente

Gráfico, Gráfico de dispersão

Descrição gerada automaticamente

Gráfico, Gráfico de dispersão

Descrição gerada automaticamente

Gráfico, Gráfico de dispersão

Descrição gerada automaticamente

**Conclusão**

Pelos resultados obtidos, chegamos a conclusão que o melhor modelo de regressão para ser utilizador seria o de Random Forest, pois apresentou um valor de r² mais próximo de 1.