**Relatório Tech Challenge**

Este relatório apresenta a análise exploratória e a modelagem de regressão para prever o valor dos custos médicos individuais cobrados por um seguro de saúde. Utilizamos um dataset sintético expandido a partir de dados originais, pré-processado para incluir colunas numéricas tratadas e categóricas convertidas.

**1 - Preparação dos Dados**

O dataset original de 1338 linhas foi expandido para 9000 linhas através de amostragem com reposição e adição de ruído em colunas numéricas ('idade', 'imc', 'encargos') para simular variabilidade e aumentar o volume de dados. As colunas foram traduzidas para português ('idade', 'sexo', 'imc', 'filhos', 'fumante', 'regiao', 'encargos'). Colunas categóricas ('sexo', 'fumante') foram transformadas em binárias (0/1) e 'regiao' foi convertida usando One-Hot Encoding.

**2 - Análise Exploratória e Correlação**

Uma análise descritiva inicial – dataset.describe() – revelou as estatísticas básicas das variáveis numéricas. A matriz de correlação (heatmap) mostrou as relações entre as variáveis. Destaca-se a forte correlação positiva entre 'fumante' e 'encargos', indicando que ser fumante é um fator significativo no custo do seguro. Outras variáveis, como 'idade' e 'imc', também mostram correlação positiva com os 'encargos', embora menos intensa que 'fumante'.

**4. Modelagem Preditiva**

Foram avaliados dois modelos de regressão para prever os 'encargos': *Regressão Linear* e *Random Forest*. Os dados foram divididos em conjuntos de treino (80%) e teste (20%).

**4.1. Modelo de Regressão Linear**

Um modelo de Regressão Linear foi treinado nos dados de treino. Os coeficientes indicam o impacto de cada variável nos 'encargos'.

**- R² (Treino):** 0.7419

**- R² (Teste):** 0.7257

**- RMSE (Teste):** 6282.76

Os P-values do modelo indicam a significância estatística de cada preditor. Variáveis com P-value abaixo de um limiar comum (exemplo: 0.05) são consideradas estatisticamente significativas para prever os encargos.

P-values e Intervalos de Confiança para os coeficientes do modelo:

**const** 7.065896e-172

**idade** 0.000000e+00

**sexo** 8.962475e-01

**imc** 2.991632e-153

**filhos** 5.754439e-12

**fumante** 0.000000e+00

**regiao\_nordeste** 2.975677e-38

**regiao\_noroeste** 1.187652e-50

**regiao\_sudeste** 1.197772e-70

**regiao\_sudoeste** 5.365895e-69

Intervalos de Confiança (95%):

|  | **0** | **1** |
| --- | --- | --- |
| **const** | -10465.914670 | -9128.705608 |
| **idade** | 240.880273 | 261.076009 |
| **sexo** | -265.200300 | 302.999572 |
| **imc** | 315.614549 | 364.988483 |
| **filhos** | 297.280091 | 533.361424 |
| **fumante** | 23576.972294 | 24292.230912 |
| **regiao\_nordeste** | -2173.773821 | -1604.400187 |
| **regiao\_noroeste** | -2498.092643 | -1923.537401 |
| **regiao\_sudeste** | -3330.774332 | -2675.493748 |
| **regiao\_sudoeste** | -2991.889652 | -2396.658494 |

A análise dos P-values e intervalos de confiança revela que a maioria das variáveis, incluindo 'idade', 'imc', 'filhos', 'sexo', 'fumante' e as dummies de 'regiao', são estatisticamente significativas para o modelo linear.

**4.2. Modelo Random Forest**

Um modelo Random Forest, que lida melhor com não linearidades e interações, foi treinado.

**- R² (Treino):** 0.9939

**- R² (Teste):** 0.9495

**- RMSE (Teste):** 2695.40

O Random Forest apresentou métricas de desempenho (R² e RMSE) superiores à Regressão Linear, indicando uma melhor capacidade de capturar a variabilidade nos dados e fazer previsões mais precisas.

A importância das variáveis no modelo Random Forest confirma que 'fumante' é, de longe, o preditor mais importante para os encargos, seguido por 'imc' e 'idade'.

**5. Análise de Resíduos (Random Forest)**

A análise dos resíduos do modelo Random Forest ajuda a avaliar a qualidade das previsões:

* A distribuição dos resíduos parece centralizada em zero, o que é desejável.
* O gráfico de resíduos vs. valores previstos mostra uma dispersão que sugere homocedasticidade (variância constante dos erros), embora possa haver um leve aumento da dispersão para valores previstos mais altos, o que é comum em dados de encargos.

**6. Insights Obtidos**

1. **Fumante é o fator dominante:** Ser fumante tem o impacto mais significativo nos encargos do seguro, superando em muito o impacto das outras variáveis.
2. **Idade e IMC são importantes:** Idade e IMC também são preditores relevantes, com encargos geralmente aumentando com a idade e o IMC.
3. **Região tem influência menor:** As dummies de região mostraram menor importância no modelo Random Forest em comparação com 'fumante', 'idade' e 'imc'. No modelo linear, algumas regiões podem ter um impacto estatisticamente significativo, mas a magnitude do efeito é menor.
4. **Random Forest supera Regressão Linear:** O modelo Random Forest obteve um R² maior e um RMSE menor no conjunto de teste, indicando que ele é mais adequado para prever os encargos neste dataset, possivelmente devido à sua capacidade de modelar interações e relações não lineares entre as variáveis.

**7. Validação Estatística**

A validação estatística foi realizada principalmente através das métricas de avaliação (R² e RMSE) nos conjuntos de treino e teste e pela análise dos P-values e intervalos de confiança no modelo de Regressão Linear.

* O RMSE mede o erro típico nas previsões. Um RMSE menor indica que as previsões do modelo estão, em média, mais próximas dos valores reais
* O R² indica a proporção da variância nos encargos que é explicada pelo modelo. Um R² mais alto (como o do Random Forest) sugere que o modelo explica uma maior parte da variância dos dados..
* Os P-values no modelo linear confirmam a significância estatística dos preditores, indicando que as relações observadas provavelmente não são devidas ao acaso.
* Os Intervalos de Confiança fornecem um intervalo dentro do qual o verdadeiro coeficiente populacional provavelmente se encontra (com um certo nível de confiança, aqui 95%).
* A análise de resíduos (para Random Forest) valida a suposição de que os erros de previsão são, em média, zero e distribuídos de forma razoavelmente aleatória, embora a homocedasticidade possa não ser perfeita.

**8. Conclusão**

Ambos os modelos de regressão foram capazes de prever os encargos de seguro com base nas características dos indivíduos. No entanto, o modelo Random Forest demonstrou um desempenho superior em termos de R² e RMSE (Regressão Linear).

A análise das importâncias das variáveis e a validação estatística confirmam a forte influência de ser fumante, idade e IMC nos custos do seguro. O relatório nos forneceu insights valiosos para entendermos os principais impulsionadores dos encargos de seguro e demonstra a eficácia dos modelos preditivos na análise destes dados.