Relatório Tech Challenge Fase 01

Welcome to IA para Devs
Relatório do Tech Challenge

Daniel Felipe Klotz - RM358514
Daniel Negreiros Cangianelli - RM359064
Felipe de Castro Sá Barreto - RM
Henrique Sartal Santos - RM358617

28/10/2024

Introdução

Objetivo do Projeto

O objetivo deste trabalho foi desenvolver um modelo preditivo para prever o custo médico individual cobrado em seguros de saúde, utilizando técnicas de regressão. Essa análise se torna relevante para compreender os fatores que mais influenciam os custos, possibilitando o desenvolvimento de políticas de preço mais eficientes e personalizadas.

Para o desenvolvimento do projeto, utilizou-se o dataset “Healthcare Insurance” do Kaggle, com variáveis como idade, gênero, IMC, número de filhos, status de fumante, região de residência e encargos médicos. Este dataset foi escolhido por conter as variáveis especificadas em nosso template de dados e apresentar um tamanho adequado para a análise.

Preparação e exploração dos dados

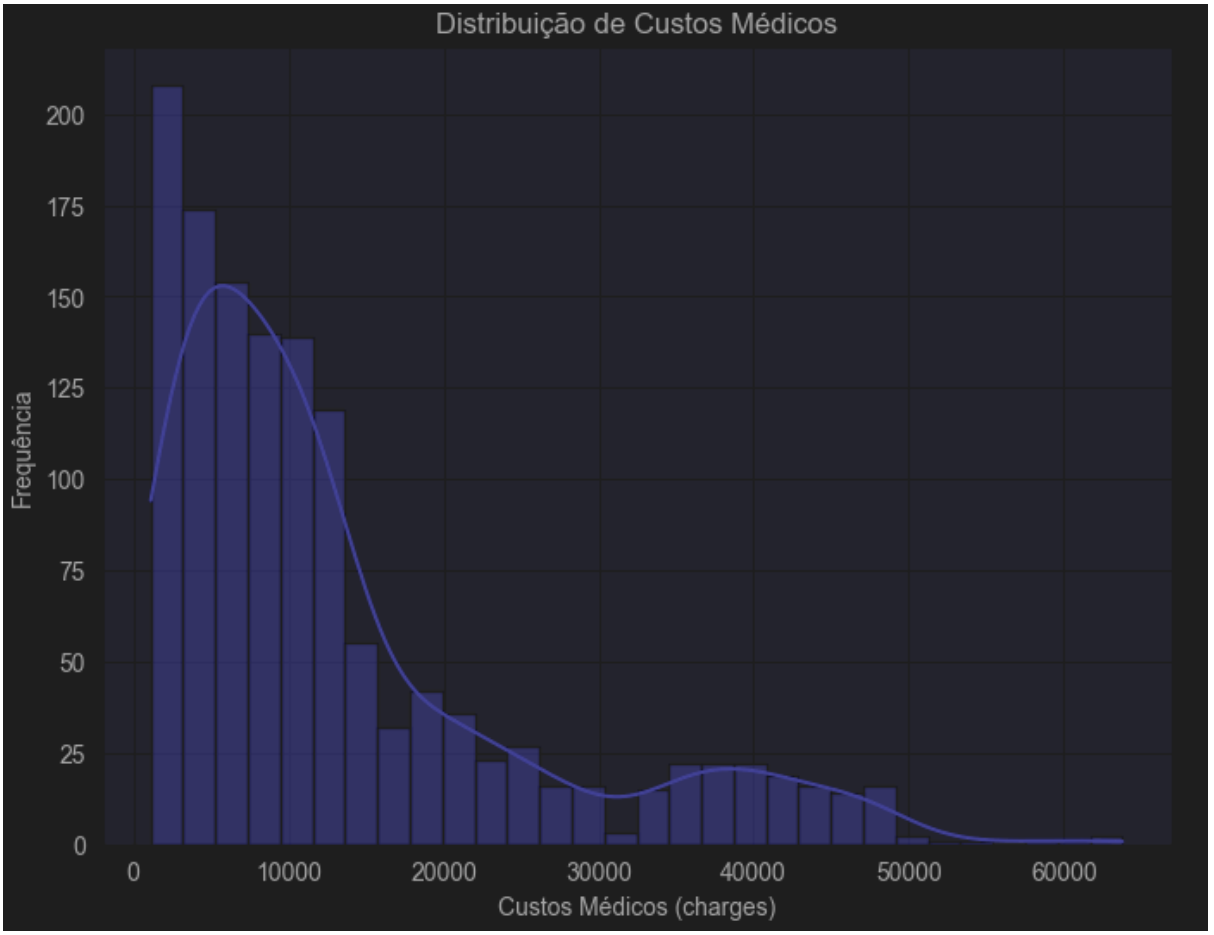
Carregamento e Verificação de Dados:

Após realizar a leitura do dataset obtido via Kaggle, foi efetuada a checagem de valores nulos e realizado tratamento de dados específicos de acordo com a variável em questão.

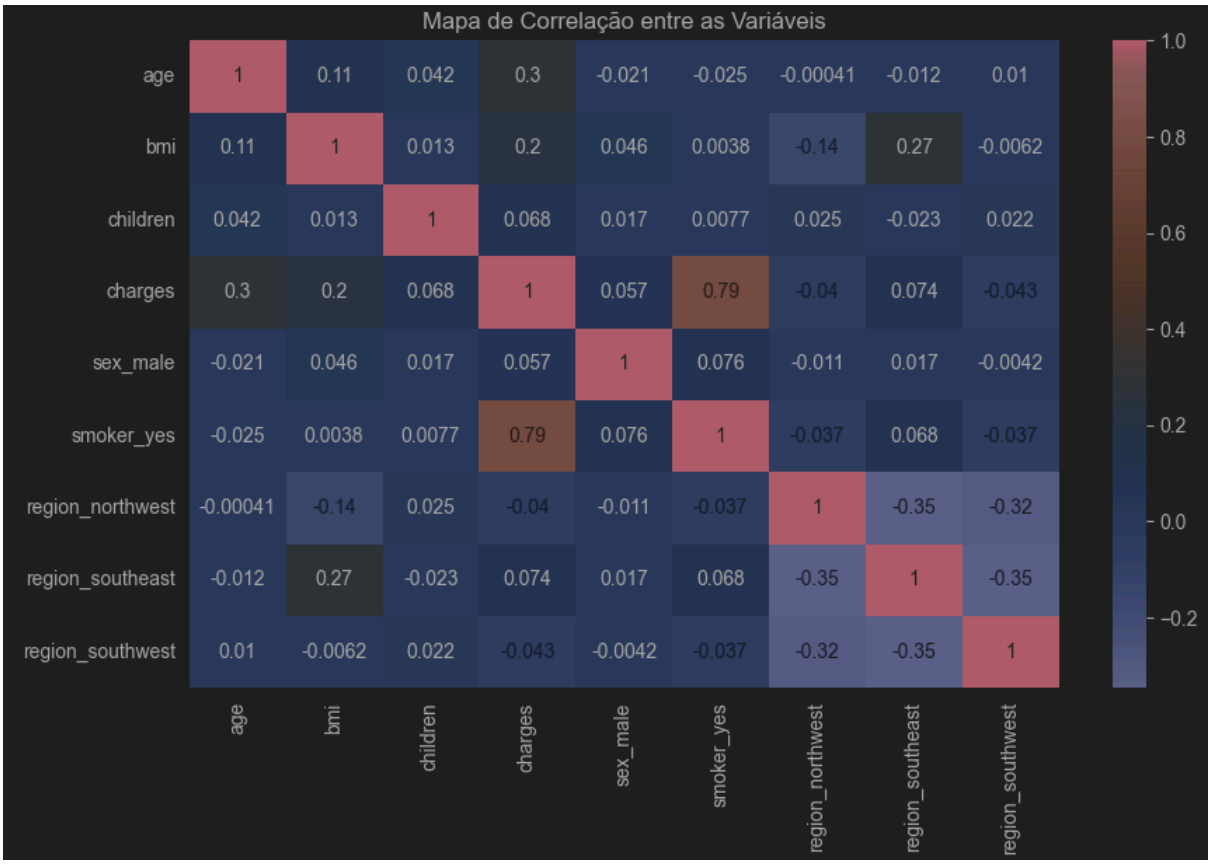
- Caso a variável fosse numérica, o valor nulo foi substituído pela mediana, uma vez que ela se mostra robusta contra outliers e oferece boa centralidade para variáveis assimétricas, ou seja, minimiza o impacto de valores extremos.
- Caso a variável fosse categórica ou fosse a variável target, removeu-se toda a linha da mesma.

Análise Exploratória:

Na análise exploratória, foi gerado um gráfico que mostra a frequência dos custos médicos. Observou-se que a maioria dos custos é baixa, mas há alguns valores bem altos. Essa informação é importante, pois modelos mais simples podem ter dificuldade para prever corretamente esses casos mais caros.



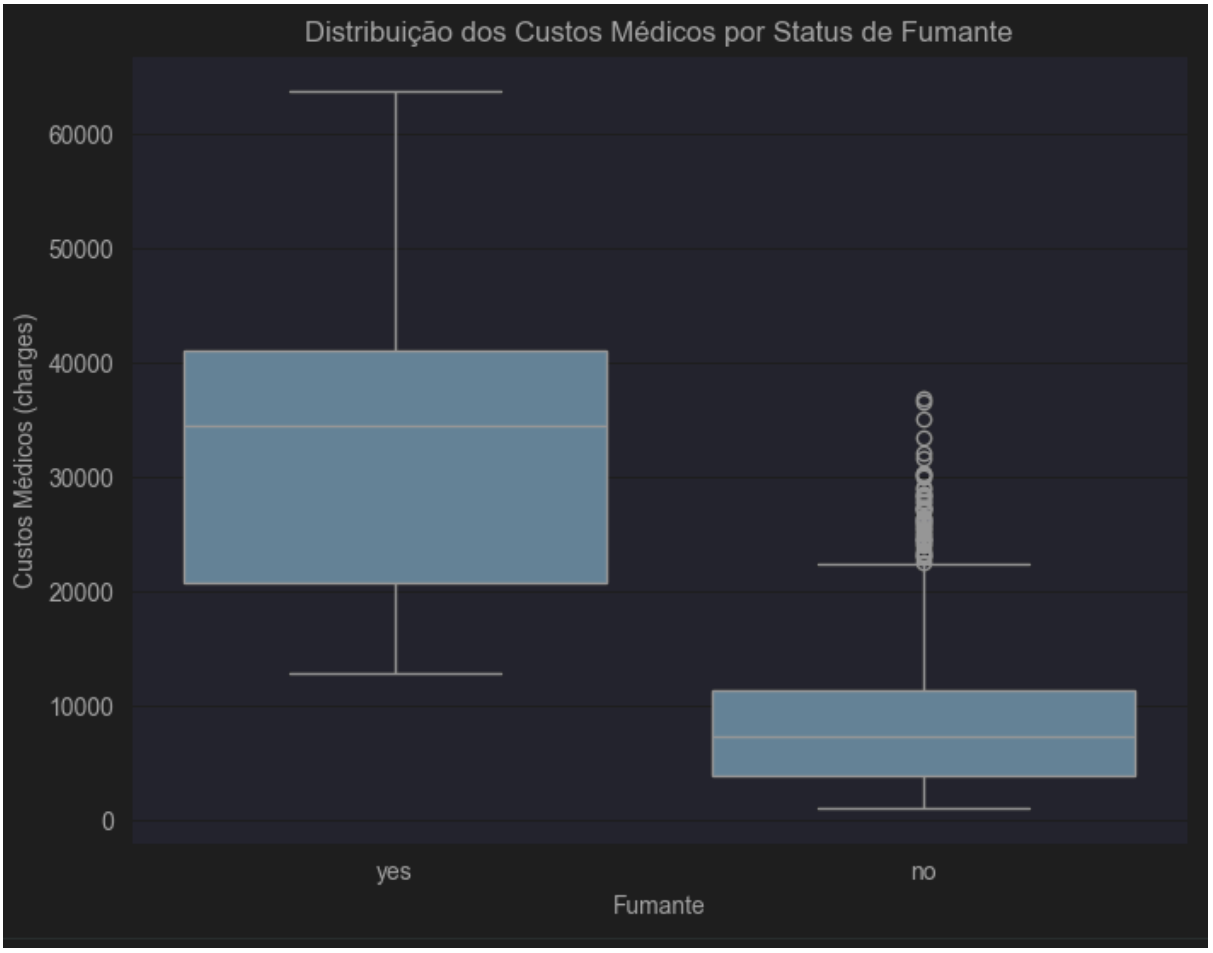
Em seguida, analisou-se a correlação entre as variáveis, usando o coeficiente de Pearson, e visualizamos a matriz de correlação com um mapa de calor. A correlação mais forte foi observada entre o status de fumante e os custos, com uma correlação de 0,79. Essa relação indica que ser fumante é um dos principais fatores que contribuem para um aumento nos encargos.



Boxplot: Comparação de Fumantes e Não Fumantes

Criamos boxplots para comparar a distribuição dos custos médicos entre fumantes e não fumantes:

- **Fumantes**: Observou-se uma mediana elevada e um intervalo amplo, sem outliers visíveis, sugerindo que os fumantes tendem a ter custos médicos elevados e consistentes.
- **Não Fumantes**: Apresentaram uma mediana baixa e maior variabilidade nos dados, com presença de outliers. Esse grupo mostra uma tendência de custos médicos menores, com casos elevados apenas em situações específicas.



Pré-processamento de Dados

Divisão de Variáveis de Entrada e Saída

Separou-se as variáveis em X (idade, gênero, IMC, filhos, fumante, região) e Y (encargos).

Padronização e Codificação das Variáveis

Para assegurar a compatibilidade dos dados com algoritmos de aprendizado de máquina:

- Variáveis Numéricas: Aplicamos a padronização, importante para modelos sensíveis à escala, como a regressão linear.
- Variáveis Categóricas: Utilizamos a técnica de One-Hot Encoding para transformar categorias em colunas binárias, evitando o tratamento ordinal inadequado de categorias.

Modelagem e Avaliação dos Modelos

Modelo 1: Regressão Linear

Treinamento e Avaliação

A regressão linear foi o primeiro modelo escolhido, visando identificar relações lineares. O modelo foi treinado e avaliado com as métricas de MAE, MSE, RMSE e R^2. Com um R^2 de 0,7832, o modelo captou parcialmente as tendências, embora tenha subestimado os custos mais altos. A análise gráfica dos valores reais versus preditos indicou que:

- Para valores mais baixos, o modelo foi relativamente preciso.
- Para valores altos, houve subestimação, sugerindo uma limitação para capturar a complexidade de casos extremos, o que pode ser resolvido com modelos não lineares.

Cross-Validation (K-Fold)

Utilizamos a validação cruzada com 5 folds, obtendo um R^2 médio de 0,74 com um desvio padrão de 0,06. Isso mostrou uma consistência razoável do modelo, indicando baixo overfitting e generalização adequada.

Modelo 2: Decision Tree Regression

Configurações e Critério de Avaliação

Optamos pelo critério de erro quadrático médio (mse) e limitamos a profundidade máxima a 5. O uso de outros critérios, como absolute_error e friedman_mse, poderia ter um impacto nos resultados:

- absolute_error: Poderia lidar melhor com outliers, pois penaliza desvios lineares, enquanto o mse amplifica grandes erros.
- friedman_mse: Focado em minimizar variâncias, poderia ser útil para dados com distribuição assimétrica.

Desempenho

A árvore de decisão obteve um R^2 de 0,8555, superando a regressão linear. No gráfico de dispersão, as previsões aproximaram-se mais dos valores reais, especialmente em casos de custo médio-alto. A árvore revelou-se mais eficaz em capturar a complexidade dos dados, sugerindo que variáveis interativas, como status de fumante, são mais bem capturadas.

Modelo 3: Random Forest Regression

Configurações e Treinamento

Utilizamos 6 estimadores e uma profundidade máxima de 6. Random Forests são eficientes em suavizar as previsões, combinando múltiplas árvores de decisão para reduzir variâncias.

Desempenho

O R^2 de 0,8635 indicou que o modelo Random Forest foi o mais eficiente em prever custos de forma consistente. A análise gráfica mostrou uma excelente adequação, especialmente para valores médios e altos, tornando o Random Forest o modelo mais robusto para capturar a variabilidade dos custos médicos.

Conclusão

Os resultados indicaram que o modelo Random Forest foi o mais eficiente, seguido pela Árvore de Decisão e, por fim, a Regressão Linear. A importância da variável Fumante foi reforçada, sendo um dos fatores que mais contribuem para a variação de custos.

Observamos que os custos mais altos foram subestimados pela regressão linear.