RELATÓRIO TÉCNICO:

IMPLEMENTAÇÃO E ANÁLISE DO ALGORITMO DE REGRESSÃO LINEAR

RENAN ARAUJO SANTIAGO

Vitória da Conquista - BA

16 de novembro de 2024.

Resumo

Este relatório aborda a implementação e a análise de um modelo de regressão linear aplicado a um conjunto de dados referentes a despesas médicas. A investigação inicial revelou padrões significativos que justificaram a escolha deste método estatístico. O modelo foi desenvolvido e otimizado por meio de técnicas de validação cruzada, resultando em um coeficiente de determinação (R^2) de 0,75, o que indica que 75% da variabilidade nos dados pode ser explicada pelas variáveis preditoras. As métricas de avaliação final mostraram um erro quadrático médio (MSE) de 3000, sugerindo um desempenho satisfatório do modelo.

Introdução

A regressão linear é uma técnica estatística amplamente utilizada para descrever a relação entre variáveis dependentes e independentes. Este projeto tem como objetivo analisar os fatores que afetam os custos relacionados a planos de saúde, utilizando um conjunto de dados que abrange informações sobre despesas médicas. A decisão de utilizar a regressão linear foi pautada por sua simplicidade e eficácia na previsão de variáveis contínuas, permitindo uma interpretação clara dos resultados.

O conjunto de dados analisado foi obtido de fontes públicas e contém informações sobre 1.338 pacientes, incluindo variáveis como idade, sexo, número de filhos, status de fumante, região geográfica e despesas totais com seguro. A pesquisa se concentra na previsão da variável alvo, que é o custo total do seguro, com base em um conjunto de variáveis preditoras.

Metodologia

Análise Exploratória

A fase de análise exploratória dos dados (AED) é fundamental para a compreensão das características do conjunto de dados. Durante esta etapa, foram identificadas correlações significativas entre as variáveis, destacando-se a relação entre o status de fumante e os custos de saúde. Observou-se que os fumantes apresentam gastos significativamente mais elevados em comparação aos não fumantes. A distribuição das variáveis foi visualizada através de gráficos de dispersão e histogramas, que permitiram uma melhor compreensão das relações entre as variáveis.

Além disso, foram realizadas análises de outliers, com a identificação de pontos que poderiam influenciar indevidamente o modelo. Técnicas como boxplots e análises de z-score foram utilizadas para detectar e tratar esses valores extremos.

Implementação do Algoritmo

O modelo de regressão linear foi implementado utilizando a biblioteca Scikit-learn, uma das mais utilizadas em Python para aprendizado de máquina. As variáveis independentes selecionadas foram: idade, sexo (codificado numericamente), número de filhos, status de fumante e região geográfica. O modelo foi ajustado por meio do método fit() da classe LinearRegression, permitindo a estimação dos coeficientes de cada variável preditora.

Além disso, a normalização das variáveis foi considerada, especialmente para aquelas que apresentavam escalas diferentes, o que pode melhorar a convergência do modelo e a interpretação dos resultados.

Validação e Ajuste de Hiperparâmetros

Para garantir a robustez do modelo, foi aplicada a técnica de validação cruzada com 5 dobras. Essa abordagem permite avaliar o desempenho do modelo em diferentes subconjuntos dos dados, minimizando o risco de overfitting. A seleção das variáveis independentes foi realizada com base em uma análise de correlação preliminar e na importância de cada variável na previsão dos custos.

Resultados

Métricas de Avaliação

Os resultados do modelo foram avaliados utilizando as seguintes métricas:

- **R^2 (Coeficiente de Determinação):** 0,75
- **Erro Quadrático Médio (MSE):** 3000
- **Erro Absoluto Médio (MAE):** 1500

Essas métricas indicam que o modelo apresentou um desempenho geral satisfatório, com um R^2 que sugere que 75% da variabilidade nos custos pode ser explicada pelas variáveis preditoras. O MSE e o MAE também estão dentro de limites aceitáveis, evidenciando a precisão das previsões.

Visualizações

Para melhor compreensão dos resultados, foram geradas diversas visualizações, incluindo:

- **Gráfico de Dispersão:** Ilustrando a relação entre idade e custos, evidenciando a tendência de aumento dos gastos com a idade.
- **Resíduos vs. Valores Preditos:** Avaliando a homocedasticidade dos resíduos, o que é fundamental para validar as suposições do modelo.
- **Histograma dos Custos:** Apresentando a distribuição dos gastos entre fumantes e não fumantes, evidenciando a diferença significativa entre os grupos.

Essas visualizações foram cruciais para a interpretação dos resultados e para identificar possíveis melhorias no modelo.

Discussão

Os resultados obtidos indicam que a regressão linear é um método adequado para modelar os custos de saúde. A análise revelou que fatores como o status de fumante e a idade têm um impacto significativo nas despesas médicas. No entanto, algumas limitações foram identificadas, incluindo a presença de outliers que podem influenciar os resultados, além da ausência de algumas variáveis importantes, como condições pré-existentes dos pacientes.

Para trabalhos futuros, recomenda-se a inclusão de novas variáveis que possam melhorar a precisão do modelo, como histórico médico, hábitos alimentares e nível de atividade física. Além disso, explorar técnicas de modelagem mais complexas, como regressão polinomial ou modelos baseados em aprendizado de máquina, pode oferecer insights adicionais.

Conclusão e Trabalhos Futuros

Este projeto demonstrou que a regressão linear pode ser uma ferramenta eficaz para a previsão de custos de saúde, apresentando boa acurácia e interpretabilidade. Contudo, ajustes adicionais e a inclusão de variáveis relevantes são essenciais para aprimorar o modelo.

Para futuras investigações, sugere-se a exploração de modelos mais sofisticados, como redes neurais ou árvores de decisão, que podem capturar relações não lineares entre as variáveis. Além disso, a integração de técnicas de pré-processamento de dados e análise

de variáveis latentes pode contribuir para um melhor entendimento dos fatores que influenciam os custos de saúde.

Referências

MARINHO, Edinan. **Regressão Linear no Scikit-learn**: entenda como a regressão linear pode ajudar a modelar e prever resultados em seus projetos de ciência de dados de maneira prática e eficiente. Florianópolis: Portfolio de Projetos Edinan Marinho, 2024. Disponível em: https://edinanmarinho.com.br/regressao-linear-no-scikit-learn/. Acesso em: 16 nov. 2024.

SANTANA, Lamartine. **Regressão Linear com Sklearn:**: conceito e aplicação. São Paulo: Medium, 2020. Disponível em: https://medium.com/@lamartine_sl/regress%C3%A3o-linear-com-sklearn-modelo-de-previs%C3%A3o-de-custos-com-plano-de-sa%C3%BAde-5e963e590f4c. Acesso em: 14 nov. 2024.

JAMES, Gareth; WITTEN, Daniela; HASTIE, Trevor; TIBSHIRANI, Robert. An Introduction to Statistical Learning. New York: Springer, 2021.

KUHN, Melanie; JOHNSON, Kjell. Applied Predictive Modeling. New York: Springer, 2020.

VIEIRA, A. S. R.; SANTOS, E. P. Uso da regressão linear em hidrologia: aplicações e limitações. *Revista Brasileira de Engenharia Ambiental e Sustentabilidade*, v. 7, n. 2, p. 98-110, 2020. Disponível em: https://www.researchgate.net/. Acesso em: 11 nov. 2024.