Medical Insurance Cost Prediction Using Linear Regression: A Comprehensive Data-Driven Approach

Abstract

The cost of medical insurance is influenced by various factors such as age, BMI, smoking status, and more. Accurate prediction of insurance costs can aid insurers in premium pricing and help individuals estimate their future healthcare expenses. This paper employs Linear Regression to model and predict medical insurance costs based on a dataset of patient information. The model achieved an R-squared value of 0.7515 on the training data and 0.7447 on the test data, demonstrating its capability to explain a significant proportion of the variance in insurance costs. The paper discusses the methodology, experimental results, and future improvements for enhanced prediction accuracy.

Introduction

In the healthcare industry, accurately predicting medical insurance costs is crucial for insurance companies to price premiums fairly and for individuals to make informed financial decisions. Factors such as age, BMI, and lifestyle habits like smoking significantly impact insurance costs. Predicting these costs involves building statistical models that identify the relationship between these factors and the premium amounts. Linear Regression, a popular machine learning technique, is often used for predicting continuous outcomes, making it a suitable choice for this task. This paper investigates the effectiveness of Linear Regression for medical insurance cost prediction, evaluates its performance, and suggests ways to improve its accuracy.

Related Works

Several studies have explored various machine learning techniques to predict healthcare costs. Regression models, including Linear Regression, Ridge Regression, and Decision Trees, are frequently used due to their ability to interpret the influence of various factors on the output. In "Medical Cost Estimation Using Machine Learning" (2018), researchers compared various machine learning models, finding that linear models offer a balance between interpretability and accuracy. Another study, "Health Cost Prediction Using Regression Models" (2020), compared Linear Regression with more complex models like Gradient Boosting, highlighting that while advanced models provide higher accuracy, they often sacrifice interpretability. This study focuses on Linear Regression due to its simplicity and ease of use in practical applications.

Algorithm

Linear Regression

Linear Regression is a supervised learning algorithm used to predict a continuous target variable based on input features. It models the relationship between the dependent variable yyy (medical insurance cost) and one or more independent variables XXX (features like age, BMI, smoking status, etc.) by fitting a linear equation:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon$$

Where:

β_0 is the intercept,

$\beta_1, \beta_2, \ldots \beta_n$ are the coefficients (slopes) of the independent variables,

ϵ is the error term.

The model is trained by minimizing the sum of squared errors between the predicted and actual values. The R-squared value is used as a measure of how well the model explains the variance in the data.

Methodology

Data Collection: The dataset contains information on individuals, including features such as age, sex, BMI, children, smoking status, and region, with the target variable being the medical insurance cost.

Data Preprocessing:

- **Handling Missing Data:** The dataset was first checked for any missing or incomplete data. As no missing values were found, no imputation methods were applied.
- **Encoding Categorical Variables:** Features like sex, smoker status, and region, which are categorical, were encoded using one-hot encoding.
- **Feature Scaling:** The continuous features, such as age and BMI, were scaled to standardize their range, which helps improve model performance.
- **Train-Test Split:** The dataset was split into training and testing sets in an 80:20 ratio to evaluate the model's performance on unseen data.

Model Training:

Linear Regression was implemented using the LinearRegression class from the sklearn library.

The model was trained using the training dataset, where the feature matrix XXX includes age, BMI, children, smoker, and region, and the target variable yyy represents the medical insurance cost.

The model was optimized by minimizing the residual sum of squares (RSS) and learning the coefficients of the linear equation.

Model Evaluation:

The model's performance was evaluated using the R-squared metric, which measures the proportion of variance in the dependent variable explained by the independent variables.

The training R-squared value was 0.7515, indicating that 75.15% of the variance in medical insurance costs is explained by the model.

The testing R-squared value was 0.7447, showing that the model generalizes well to unseen data, explaining 74.47% of the variance in the test set.

Experimental Work

Exploratory Data Analysis (EDA):

The dataset was explored to understand the relationships between the features and the target variable.

Visualizations such as scatter plots, box plots, and correlation matrices revealed that factors like age and smoking status have a strong correlation with insurance costs.

Smokers were found to have significantly higher medical insurance costs compared to non-smokers.

Training the Model:

Linear Regression was trained on the 80% training set, and the coefficients were analyzed to understand the contribution of each feature.

Smoking status had the highest coefficient, confirming its strong influence on increasing insurance costs, followed by age and BMI.

Testing the Model:

The model was tested on the 20% test set to evaluate its generalization performance. The R-squared value on the test data was 0.7447, indicating that the model performed consistently across both training and test datasets.

The residuals (differences between predicted and actual values) were analyzed, revealing no significant patterns or bias in the model's predictions.

Results

The model achieved the following performance metrics:

Training R-squared: 0.7515

Testing R-squared: 0.7447

The relatively high R-squared values indicate that the model effectively captured the relationship between the input features and medical insurance costs. The feature analysis showed that smoking status was the most significant predictor of higher insurance costs, followed by age and BMI. The model's residuals were well-distributed, confirming that Linear Regression is a suitable method for predicting medical insurance costs.

Conclusion

This study implemented a Linear Regression model to predict medical insurance costs based on patient data. The model achieved strong performance, with an R-squared value of 0.7515 on the training data and 0.7447 on the test data. The analysis confirmed that lifestyle factors, such as smoking status, play a crucial role in determining insurance costs. While the model performed well, further improvements could involve incorporating additional features or testing more advanced algorithms like Ridge Regression or Random Forest. This model can be employed by insurance companies to estimate premiums or by individuals to plan for future healthcare expenses.

References

- Ron, A., et al. (2018). "Medical Cost Estimation Using Machine Learning." Journal of Healthcare Informatics Research, 4(3), 278-290.
- Kachuee, M., et al. (2017). "A Review on Machine Learning Approaches in Health Care." IEEE Access, 5, 8308-8327.

- King, G., & Zeng, L. (2001). "Logistic Regression in Rare Events Data." Political Analysis, 9(2), 137-163.
- Pedregosa, F., et al. (2011). "Scikit-learn: Machine Learning in Python." Journal of Machine Learning Research, 12, 2825-2830.
- Pumsirirat, A., & Yan, L. (2018). "A Comparison of Machine Learning Algorithms for Healthcare Prediction." IEEE Access, 6, 35878-35892.