Report on Linear Regression using Scikit-learn

Gautam Agarwal (24n0463)

Introduction

This report summarizes the implementation of Linear Regression using Scikit-learn. The task involves predicting medical charges based on patient attributes such as age, BMI, smoking status, and region. The notebook explores linear regression with one and multiple variables, evaluates performance, and analyzes feature importance.

Dataset Description

The dataset is obtained from an online repository and contains information about individuals' demographics, lifestyle, and medical charges. The key attributes include:
- Age
- Sex
- BMI
- Number of children
- Smoking status
- Region
- Medical charges (target variable).

Exploratory Data Analysis

The dataset was explored using descriptive statistics and visualizations. Key observations include:
- Smokers tend to have significantly higher medical charges.
- BMI and age show a positive correlation with charges.
- Categorical variables such as sex and region require encoding before applying regression models.

Linear Regression with One Variable

A simple linear regression model was trained using a single predictor variable (e.g., BMI or age). The fitted regression line demonstrates how charges increase with higher values of the predictor. The slope represents the change in charges for a unit increase in the predictor, while the intercept represents baseline charges.

Linear Regression with Multiple Variables

Multiple regression was performed using all features. Categorical variables were converted to numerical format using one-hot encoding. The model was trained and evaluated using R^2, Mean Absolute Error (MAE), and Mean Squared Error (MSE). The results showed that multiple regression improves prediction accuracy compared to single-variable regression.

Regression Coefficients & Feature Importance

The regression coefficients reveal the contribution of each feature towards predicting medical charges. The most significant factor influencing charges is smoking status, followed by BMI and age. This indicates that lifestyle choices such as smoking substantially impact healthcare costs.

Alternative Models & Comparisons

Although the notebook primarily focused on linear regression, it also mentions alternative models such as decision trees and random forests. These models can capture nonlinear relationships and may provide better accuracy in certain cases.

Conclusions

The analysis highlights that smoking, BMI, and age are the strongest predictors of medical charges. Linear regression provides an interpretable and effective method for prediction but may struggle with complex nonlinear patterns. Future improvements could involve regularization techniques (Ridge/Lasso regression) or advanced ensemble models.