Precision Medicine: Predicting Hospital Visit Costs with Ridge Regression

Abstract:

The rising costs associated with hospital visits pose significant challenges for healthcare providers and patients alike. In this research paper, we present a data-driven approach aimed at accurately predicting the cost of hospital visits for individual patients. Leveraging demographic factors such as age, body mass index (BMI), number of children, sex, and smoking status, we developed a predictive model using Ridge regression. Through rigorous analysis and iterative refinement, we demonstrate the effectiveness of our approach in improving cost prediction accuracy and discuss potential avenues for further enhancement.

Introduction:

The ability to accurately estimate hospital visit costs plays a pivotal role in healthcare economics, influencing resource allocation, financial planning, and patient decision-making. Existing cost prediction models often lack precision due to complex interactions between demographic variables and healthcare expenses. In response to this challenge, we embarked on a project to develop a more robust and accurate predictive model using machine learning techniques.

Methodology:

Our methodology involved several key steps, starting with data collection and preprocessing. We gathered demographic data from hospital records, including age, BMI, number of children, sex, and smoking status. To address multicollinearity and overfitting issues, we employed Ridge regression, a regularization technique that imposes a penalty on regression coefficients. Furthermore, we devised a preprocessing pipeline to encode categorical variables and generate interaction terms for quantitative features. Categorical variables were transformed using the OneHotEncoder transformer, while interaction terms were created using the PolynomialFeatures transformer. This pipeline ensured that the dataset was appropriately prepared for modeling, capturing complex relationships among variables.

Results:

Our initial model using linear regression yielded unsatisfactory results, with a high root mean square error (RMSE) and a correlation coefficient below our target threshold. However, upon implementing Ridge regression and refining our preprocessing pipeline, we observed a significant improvement in model performance. The correlation coefficient increased to 0.74, and the RMSE decreased to $5,830, indicating a closer alignment between predicted and actual hospital visit costs.

Discussion:

The success of our model highlights the effectiveness of Ridge regression in addressing multicollinearity and overfitting, thereby improving predictive accuracy. Furthermore, our preprocessing pipeline, encompassing categorical encoding and feature engineering, proved instrumental in capturing complex relationships within the data.

Moving forward, several opportunities for improvement and further research exist. Expanding the dataset to include additional variables, such as pre-existing medical conditions and socioeconomic status, could enhance predictive capabilities. Additionally, exploring alternative machine learning algorithms and validation techniques may yield insights into optimizing model performance and generalizability.

Conclusion:

In conclusion, our research demonstrates the potential of data-driven approaches in enhancing hospital cost prediction. By leveraging demographic factors and machine learning techniques, we developed a predictive model capable of accurately estimating hospital visit costs. As we continue to refine our model and explore new avenues for improvement, we aim to contribute to the advancement of healthcare economics and ultimately improve patient care.