Literature Review

Supervised Learning for Healthcare Cost Prediction

Goal: The study compares different supervised learning models to see which one predicts healthcare costs the best. It primarily focuses on accuracy across different cost groups to improve cost estimation.

Dataset: The study uses health insurance data from University of Utah Heal Pans from 2013-2016. It started with 6.3 million claims and 1.2 million pharmacy claims, but after filtering, it includes 3.8 million medical records and 780K pharmacy claims from 24K patients. The data is split into an observation period between 2013-2015 and a results period of 2015-2016.

Methodology: Looked at existing research on cost prediction models, extracted cost related features, and grouped patients into 5 cost categories. Tested multiple supervised learning models Linear Regression, Lasso, Ridge, CART, Elastic Net, M5, Bagging, Random Forest, Gradient Boosting, SVM and ANN. These models were optimized using a high-cost cross validation, and their statistical significance was checked using Friedman's test and Wilcoxon Signed-Rank test.

Results: Gradient Boosting showed to have performed the best for low to medium cost patients, while ANN and Ridge Regression were the most accurate for high-cost cases. Metrics used include Mean Absolute Percentage Error (MAPE), R^2, showing the best overall accuracy. The study's main limitations were relying on a single dataset and only using cost-related features.

Article Link: https://pmc.ncbi.nlm.nih.gov/articles/PMC5977561/

Analyzing Healthcare Expenditures Using Machine Learning

Goal: This study looks at whether last years healthcare costs can predict current costs using ML. It examines the relationship between medical and dental healthcare expenditures and aims to determine if dental costs have a measurable impact on future medical expenses.

Dataset: Utilizes Japan's area-based public health insurance data base from Ebina City in Kanagawa Prefecture. The dataset includes 30,340 inviduals continuously insured from April 2017 to September 2018, covering total healthcare expenditures of over 13.5 billing Japanese Yen (JPY).

Methodology: It applies a few different predative models, including Generalized Linear Models, Zero-Inflated Models, Support Vector Machine Regression, Neural Networks, and Generalized Boosted Regression Models. Independent variables include age, se, previous year's medical expenses, and dental expenses.

Results: It's found that last year's medical costs were the strongest predictor of current healthcare costs. Dental costs had a slight but catechistically insignificant reducing effect on future medical expenses. Of the tested models, neural networks performed for healthcare cost prediction.

Article Link: https://www.mdpi.com/1660-4601/18/2/565

Predicting Visit Costs for Obstructive Sleep Apnea Patients

Goal: This study focuses on predicting the visit costs of patients with Obstructive Sleep Apnea (OSA) using electronic healthcare records (EHR). By leveraging deep learning techniques, particularly Transformer models, the study aims to improve cost prediction accuracy and aid healthcare decision maker in budget allocation and resource planning.

Dataset: The dataset consists of EHRs from Turku University hospital in Finland, spanning from 2002 to 2019. The study focuses on patients diagnosed with OSA, filtering records to include only those with at least one year of follow-up data.

Methodology: The research employes a two-Transformer model approach-one transformer is used for augmenting the input data by incorporating information from patients with shorter visit histories, while the second Transformer predicts costs based on long-term data. The model applies deep learning techniques such as multi-task loss functions and sequential forecasting.

Results: The two-Transformer model significantly improved cost prediction accuracy, with an R^2 score increasing from 88.8% of single-Transformer to 97.5% of two-model approach. Even using the augmented data with simpler baseline models improved R^2 from 61.6% to 81.9%. These results suggested that the proposed approach makes better use of limited high-quality data, providing more accurate cost estimates.

Article Link: https://ieeexplore.ieee.org/document/10128115

Healthcare Waste Cost Prediction in Public Hospitals

Goal: This study focuses on identifying and predicting healthcare waste management costs to create an optimal and sustainable waste management system using data from Greek public hospitals.

Dataset: The study uses data from all Greek public hospitals, including management costs, waste quantities, and hospital characteristic like inpatient numbers, bed count, and surgeries performed.

Methodology: Regressions analysis was used to assess factors affection healthcare waste management costs and quantities. A few models, including ordinary least squares, seemingly untreated regression models, and multivariate regression analysis were applied to determine cost drivers.

Results: The study found that healthcare waste costs are significantly influenced by factors like numbers of beds, hospital type, ICU, number of inpatients, and hospital staff size. Two prediction models were developed to estimate waste management costs and quantities for Greek public hospitals, achieving R^2 above 85% indicating strong predictive accuracy.

Article Link: https://www.mdpi.com/1660-4601/19/16/9821

Leveraging Machine Learning for Healthcare Cost Forecasting

Goal: The study looks at the potential of deep learning models to predict future healthcare expenses using chest radiographs. The objective is to assess whether image-based data can classify top-50% healthcare spenders and provide reliable cost predictions over 1-, 3-, and 5-year periods.

Dataset: The dataset includes 21,872 frontal chest radiographs from 19,524 patients with at least one year of spending data. Subsets include 11,003 patients with 3 year cost data and 1,678 patients with 5 year cost data. Spending data covers direct and indirect expenses such as inpatient stays, pharmacy, imaging, and medical consultations.

Methodology: The study applied deep learning models, ResNet based architecture for regression and classification. Four models were tested: a baseline model using demographic data (age, sex, ZIP code), a model using only chest radiographs, a hybrid approach combining both, and an end-to-end deep learning model. Performance metrics include ROC-AUC for classification and Spearman's correlation for cost prediction.

Results: Note worthy model for predicting 1 year expenses achieved an ROC-AUC of 0.806 and a Spearman's correlation of 0.561. For a 3 year prediction, the highest ROC-AUC was 0.771, and for a 5 year prediction, 0.729. The study found that chest radiographs contain valuable indicators for future costs, suggesting an innovative approach to predictive healthcare analysis.

Article Link: https://www.nature.com/articles/s41598-022-12551-4

Hybrid Modeling for Cost Prediction with CGBN

Goal: The study explores hybrid machine learning algorithms for healthcare cost prediction. The goal is to improve cost prediction accuracy by combining Conditional Gaussian Bayesian Networks (CGBN) with regression models, reducing the reliance on expert-selected features while optimizing predictive performance.

Dataset: Two public healthcare datasets from Kaggle were used, one with 986 instances (10 features + healthcare cost) and another with 1,338 instances (6 features + healthcare cost). Both datasets contain a mix of categorical and numerical variables linked to patient demographics, medical history, and healthcare expenses.

Methodology: The research applies a hybrid approach where CGBN is used to filter out irrelevant features from the datasets before training regression models. Several regression techniques were tested, including Linear Regression, Support Vector Regression (SVR), Backpropagation Neural Networks (BPnet), Random Forest (RF), and Long Short-Term Memory (LSTM). The models were trained and evaluated using RStudio with 80% of the dataset allocated for training and 20% for testing.

Results: The hybrid models showed improved prediction performance while reducing dataset complexity. CGBN combined with RF realized the lowest Mean Relative Error (MRE), outperforming standalone models. Other metrics, including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Symmetric Mean Absolute Percentage Error (SMAPE), also shown the effectiveness of hybrid modeling. The study found that using CGBN to filter out unnecessary data helped models make better predictors without needing large datasets.

Article Link: https://www.mdpi.com/2227-7390/11/23/4778