Enhancing Healthcare Expenditure Prediction in Diabetes: A Machine Learning Approach

MSR221

Hyung-Seok (John) Kim MHS[1], Yu-Hua Fu MS[1], Pei-Lin Huang, MHS[1], Zafar Zafari PhD[1,2]
[1]University of Maryland School of Pharmacy, Baltimore, MD, USA, [2]The University of Maryland School of Pharmacy and Institute for Health Computing, North Bethesda, MD, USA

UNIVERSITY of MARYLAND SCHOOL OF PHARMACY

Introductions

Diabetes imposes a substantial financial burden on society with $1 out of every $4 in United States (US) healthcare costs allocated to its management.

Despite this, few studies have explored the potential of machine learning (ML) algorithms to improve predictions of healthcare expenditure.

Objective

To develop ML algorithms to predict total healthcare expenditures among diabetics.

Methods

- **Study Design:**
  - Cross-sectional study
- **Data source:**
  - Full-year consolidated data from the 2021 Medical Expenditure Panel Survey (MEPS)
- **Inclusion criteria**
  - Individuals aged 18 years and above with diabetes
- **Features:**
  - Thirty variables were considered.
  - Demographics: sex, age, race, ethnicity, socioeconomic status, region, marital status, education, employment, insurance status, affordability for basics
  - Comorbidities: smoking status, high blood pressure, coronary heart disease, angina, myocardial infraction, other heart disease, stroke, emphysema, high cholesterol, cancer, arthritis, asthma, diabetes-related complications kidney problem and eye problem
  - Diabetes treatments: diet modification, oral anti-diabetic medication, insulin
- **Study outcome:**
  - Total healthcare expenditures were log-transformed and adjusted to 2023 US dollars using the consumer price index.

Methods (Cont.)

- . **Model Building and Performance:**
  - The predictive performances of traditional regression, lasso, random forests (RF), tree-based boosting, neural networks (NN), and stacked ensemble methods were compared with 10-fold cross-validation.
  - The data were split into a 3:7 ratio.
  - Performance metrics included mean square error (MSE) and correlation between the models' predicted outcomes and observed outcomes.
  - Model complexity was assessed through computation time.

Results

The study included 3,184 individuals.

Linear regression yielded an MSE of 0.60 and a correlation of 0.63. Lasso, with a similar computation time, yielded an MSE of 0.58 and a correlation of 0.61.

ML algorithms showed similar performances to that of linear regression. Stacked ensemble models did not improve performance but required 10 times the computation time of linear regression's. Tree-based boosting, RF, and NN performed the best with MSE (0.53 – 0.55) and correlations (0.65 – 0.68) at the expense of nearly 7 to 20 times computation time.

Table 1. Comparisons of predictive performance across 7 models.

Predictive Performances of Models

Methods	Parameters	MSE	Correlation
Random Forests	200 trees + 11 features	0.53	0.65
Neural Networks	1 layer + 15 neurons	0.54	0.66
Tree Based Boosting	709 trees + tree depth 7	0.55	0.68
Lasso	Lambda 0.01 + 33 non-0 features	0.58	0.61
Stacked Ensemble: Gradient Boosting	Components: RF + NN + LM	0.60	0.60
Linear Regression	-	0.60	0.63
Stacked Ensemble: Regression	Components: RF + NN + LM	0.61	0.60

MSE = mean squared error, Correlation = Pearson correlation

Results (Cont.)

Figure 1. Comparison of model performance (MSE) on computation time.

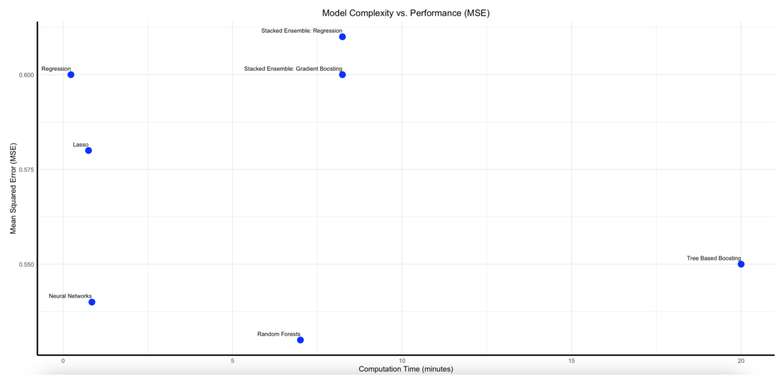
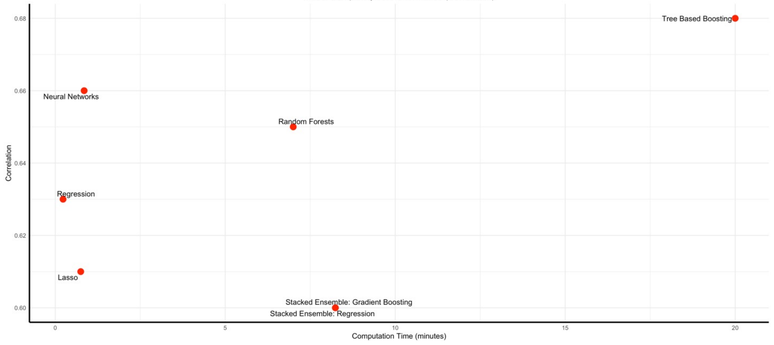


Figure 2. Comparison of model performance (correlation) on computation time.



Conclusion

In the context of our study, ML algorithms offered minimal improvements over linear regression at the expense of substantially increased computation time. ML models can effectively predict outcomes but may be more appropriate in scenarios where capturing complex non-linear relationships between variables is paramount.

Contact Information

Hyung-Seok (John) Kim , MHS

PhD student, Department of Practice, Sciences, and Health Outcomes Research, University of Maryland, Baltimore

E-mail: hkim4@umaryland.edu