https://doi.org/10.24867/SYMOPIS-2024-51-096

OPPORTUNITIES FOR HEALTHCARE COST PREDICTION USING MACHINE LEARNING ALGORITHMS

TATJANA RAKONJAC–ANTIĆ[1], MARIJA KOPRIVICA[2], MILICA KOČOVIĆ DE SANTO[3],KRISTINA BRADIĆ[4]

[1] University of Belgrade – Faculty of Economics and Business, Belgrade, tatjana.rakonjac@ekof.bg.ac.rs, ORCID: 0000-0003-0371-0115
[2] University of Belgrade – Faculty of Economics and Business, Belgrade, marija.koprivica@ekof.bg.ac.rs, ORCID: 0000-0003-4239-2252
[3] Institute of Economics Sciences, Belgrade, milica.kocovic@ien.bg.ac.rs, ORCID: 0000-0003-3304-7801
[4] Master's student – Faculty of Economics and Business, Belgrade, bradickristina1@gmail.com

Abstract: *The growing trend of healthcare costs, increased life expectancy, and the increasing availability of data on policyholders indicate the importance of the application of machine learning in health insurance. Using historical data of policyholders, machine learning enables the prediction of healthcare costs, identification of high-risk individuals for hospitalisation, assessment of the likelihood of chronic diseases, and more. The subject of research in this paper are the opportunities for healthcare cost prediction by implementing different machine learning algorithms. Based on the public database from the Kaggle website, the created model incorporates various machine learning algorithms such as Random Forest, Gradient Boosting and Linear Regression. The aim of the paper is to point out that selecting a predictive machine learning model with the best performance can significantly improve the prediction of individual healthcare costs. This, in turn, contributes to determining appropriate premiums for voluntary health insurance.*

Keywords: *healthcare costs, health insurance, insurance premium, machine learning, database, Random Forest, Gradient Boosting, Linear Regression.*

1. INTRODUCTION

The healthcare system, by performing the function of treating and improving the health condition of the population, plays a crucial role in the economic development of a country. However, one of the biggest challenges governments today face is the continuous rise in healthcare costs. The importance of funding and effective coordination of the healthcare system has been particularly highlighted during the COVID-19 pandemic, which caused an unforeseen increase in healthcare expenditures [7].

In health care, big data is generated by various sources such as wearable medical devices, health apps, electronic health records (EHR), and electronic medical records (EMRs), enabling a wide range of new approaches to accurately predicting healthcare costs. Factors like health status, demographic information, geographic access, and lifestyle choices play a crucial role in determining potential healthcare costs. As the volume of data increases, manual calculation becomes sluggish and error-prone. In such conditions, implementing machine learning models for predicting healthcare costs can be highly beneficial for insurance companies. Machine learning models iteratively learn from past data and facilitate predicting claims for healthcare services submitted by policyholders [3, 7, 9]. The subject of our research is the prediction of healthcare costs by implementing different machine learning algorithms. In the research, we used dataset from the Kaggle repository.

2. LITERATURE REVIEW

Over the past decade, insurance companies have increasingly turned to the implementation of machine learning and artificial intelligence to enhance risk assessment. Additionally, numerous researchers aim to determine healthcare costs using various machine learning algorithms in different contexts. The quality of data collection, data preprocessing, and feature selection determines the model's performance [9].

The authors utilized medical insurance cost dataset from the Kaggle repository to provide a computational intelligence approach for predicting healthcare costs. The research proved that the Stochastic Gradient Boosting (SGB) model outperformed the others (Linear Regression, Support Vector Regression, Ridge Regressor, Stochastic Gradient Boosting, XGBoost, Decision Tree, Random Forest Regressor, Multiple Linear Regression, and k-Nearest Neighbors), achieving 86% accuracy and a cross-validation value of 0.0858 [14].

Recent medical insurance research has increased interest in predicting costs at the level of the individual member, with emphasis on high-cost claimants (HiCCs) - patients whose annual healthcare expenditures exceed $250,000, constituting 9% of total healthcare costs in the United States. Maisog et al. (2019) analysed health insurance claims from 48 million individuals and census data to develop binary classification models to calculate the personal risk of HiCC. The study proved that leveraging claims data and publicly available information allow development of high-performing predictive models. The Light Gradient Boosting (LGBM) model, in particular, had the highest Area Under the Receiver Operating Characteristic Curve (AUC-ROC) score at 91.25% [6].

To predict healthcare costs, Panay et al. (2019) utilized 25,464 Japanese health records from Tsuyama Chuo Hospital's, encompassing billing information, medical checkups, exam results spanning from 2016 to 2017. Their research found that Interpretable Evidence Regression (IEVREG) achieved the highest performance in healthcare cost prediction with R^2 of 0.44 in comparison with Gradient Boosting and Artificial Neural Network [8].

Vimont et al. (2022) used a 1/97 representative sample from the French National Health Data Information System which contained 510,182 patient records, including patient's demographic information, pre-existing conditions, Charlson comorbidity index, healthcare service use and costs. The results showed that the Random Forest model should be preferred for predicting healthcare costs based on individual risk adjustment, reaching an adjusted R^2 of 34.7% and Mean Absolute Error (MAE) of EUR 1338 [15].

Kshirsagar et al. (2021) developed a sequence of two models (an employer-group-level model and an individual patient-level model) using pharmacy, medical and capitation claims data from 14 million patients. Their research evaluated the ability of the machine learning algorithms to predict the per member per month cost for employer group in the next renewal period, focusing on the groups with potential cost-saving opportunities. The models identified 84% of these opportunities and performed 20% better than the insurance carrier's existing pricing model [5].

Finally, Sahai et al. (2023) performed a comparative analysis between tree-based classifiers such as XGBoost, Decision Tree and Random Forest to propose the most accurate model for insurance risk classification and prediction. The authors highlighted the importance of model interpretability for stakeholders, financial institutions and regulators. Their findings demonstrated that the XGBoost classifier performed the best when compared to others with AUC value of 0.86 [11].

3. METHODOLOGY AND DATA

In this study, we used publicly available data from the Kaggle repository. The dataset, consisting of 2,772 records, underwent initial data preprocessing and the selection of relevant variables for inclusion in the model. Subsequently, the prepared dataset was divided into two parts, one used for iterative model training and the other for testing the models. The training set was used to create

models predicting annual healthcare costs, while the test set evaluated model performance. The methodology steps are presented in Figure 1.

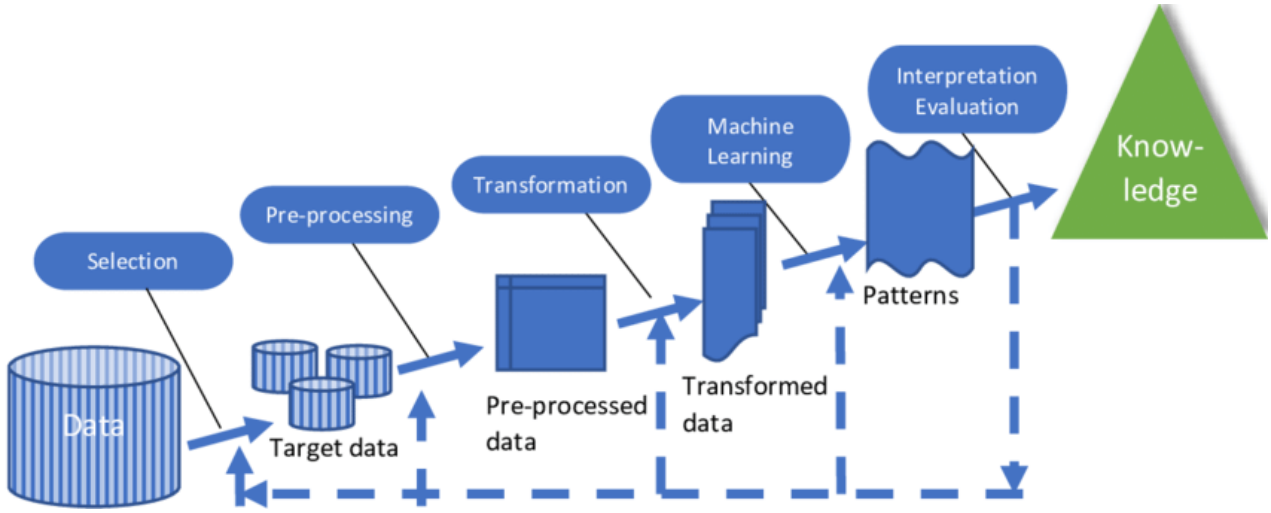

Figure 1: Applied Methodology
Source: [4]

Within the acquired database, we identified 7 attributes as described in Table 1. Checks for missing values were done. Given the different types of data across the attributes, we employed the Label Encoder technique to transform categorical variables into integers (0, 1, 2, 3), depending on the number of unique values within each variable. Exploratory data analysis (EDA) was conducted to uncover the correlation between some key features by the Pearson's correlation heatmap. Smoking habits, body mass index, and age were identified as attributes exerting the greatest influence on healthcare costs. Moreover, the analysis revealed that healthcare costs incurred by smokers residing in specific regions exert a considerable influence on insurance premiums. This relationship underscores the regional disparities in healthcare expenditures attributable to smoking-related illnesses, thereby affecting the pricing strategies adopted by insurers.

Table 1: Dataset description

Variables	Description	Data type
age	Age of policyholder	Numeric
sex	Gender of policyholder	Categorical
bmi	Body mass index, weight/[height]^2	Numeric
children	Number of children policyholder have	Numeric
smoker	Smoking habit of policyholder	Categorical
region	Region where policyholder lives	Categorical
charges	Insurance premium	Numeric

In the course of the research, we developed three machine learning models based on the ensemble model concept, using Random Forest and Gradient Boosting, along with the Linear Regression algorithm.

Linear Regression belongs to the group of the supervised machine learning algorithms whose aim is to predict the value of the dependent variable (y) based on independent variables (x) [12]. The primary advantage of this algorithm lies in its simplicity and ease of application, especially in cases where a linear relationship between the dependent and independent variables is known. However, in most real-world scenarios, the relationship among variables is complex and far from linear [10].

Gradient Boosting is an iterative method aimed at minimizing the loss function, i.e., the error between predicted outputs and actual targets. During each phase of training, the error is computed,

and thus, the parameters of the function are updated. The algorithm's advantage lies in continuously improving the model, although it is computationally exhaustive [13].

Lastly, Random Forest utilizes a set of decision trees generated from the dataset, where data is distributed among trees using various ensemble learning techniques. The model aggregates the predictions from multiple trees to compute the final results, typically by averaging the outputs of randomly selected trees [1].

Table 2: Comparative analysis of the R-squared results for the models

	Linear Regression	Random Forest	Gradient Boosting
R^2	0.771	0.951	0.956

In Table 2, we present a comparative analysis of the R-squared results for three models, concluding that the Gradient Boosting model exhibited superior performance with a coefficient of determination of 0.956. This high value suggests that 95.6% of the variability in the dependent variable can be explained by the independent variables incorporated into the Gradient Boosting model.

4. APPLICATION POSSIBILITIES OF MACHINE LEARNING IN A LOCAL HOSPITAL

Building on the methodology adopted from the Kaggle database, we aimed to leverage machine learning to the operational cost database of a local hospital in Serbia. Upon analyzing data from March 2021, 2022, and 2023, it became evident that healthcare costs surpassed revenues by RSD 31,763, 31,413, and 35,290, respectively. This underscores the critical need for implementing effective tools aimed at enhancing the efficiency of healthcare system funding and coordination.

The acquired database of costs from the local hospital did not enable the implementation of machine learning due to the lack of detailed documentation regarding the sources of these expenditures and patient information. The database was compiled through manual entry of costs per department (cardiology, surgery, gynecology and obstetrics, pediatrics, extended care) without disclosing the underlying factors that determined these costs. This lack of information has prevented the development of machine learning models based on input and output variables. A database conducive to the successful application of machine learning algorithms in predicting patient healthcare costs would encompass demographic and administrative patient data (e.g., age, sex, place of residence), records of healthcare visits and reimbursed procedures (e.g., medications, medical procedures, medical devices, laboratory tests). Aggregating the dataset with information from state and private hospitals would enrich the overall repository, capturing additional patient-specific details such as medical histories, diagnoses, external procedures, costly medications, and implantable devices.

By analyzing the revenue structure of considered hospital for March 2023, we concluded that only 1% of the revenue (RSD 889,949) comes from collected copayments and insurance premiums, while over 95% of the revenue stems from contracts with the Republic Health Insurance Fund and E-invoices. Such a revenue structure presents ample opportunities for its enhancement, primarily through diversification of revenue sources. A larger share of contracts with insurance companies would contribute to higher total revenues and more efficient healthcare system management. Following the practice of the Australian healthcare system, voluntary health insurance could enable healthcare in both private and public hospitals [2]. This approach would ensure patient health records are accessible to any healthcare institution and selected insurance company, regardless of the treatment location, significantly improving healthcare quality and the insurance risk assessment model. By implementing machine learning algorithms on such databases, insurance companies would have the opportunity to optimize insurance premium determination models.

It is important to emphasize that improving healthcare cost management begins with reforming patient and business processes data collection. The successful implementation of machine learning

relies on having a standardized database that encompasses all variables that contribute to the final patient bill within the framework of the national healthcare system. The opportunities for predictive modeling continue to grow and improve, but they are directly correlated with the data collection, processing, and analysis. A robust approach to implementing an electronic health record system provides the opportunity to leverage all the benefits of machine learning. Otherwise, rising healthcare costs and increased life expectancy will continue to place further pressure on the healthcare system, spilling over into the entire economy.

5. CONCLUSION

Big Data generated from sources such as wearable medical devices, health apps, electronic health records (EHR), and electronic medical records (EMRs) enables a wide range of new approaches in accurately predicting healthcare costs. Factors like health status, demographic information, geographic access, and lifestyle choices play a crucial role in determining potential healthcare costs. In such conditions, implementing machine learning models for predicting healthcare costs can be highly beneficial for the insurance companies but also policyholders.

In this paper, three machine learning models, Linear Regression, Random Forest, and Gradient Boosting were deployed for predicting healthcare costs using medical insurance cost dataset from Kaggle repository. The Gradient Boosting model achieved the highest R-squared score of 0.956, proving that the selection of the best performance model may significantly contribute to the improvement of the individual healthcare cost prediction. This improvement can lead to more accurate determination of voluntary health insurance premiums. However, our attempts to apply machine learning to a local hospital's operational cost database were hindered by inadequate documentation of expenditure sources and patient information. Effective machine learning implementation hinges on a comprehensive, standardized database that encompasses all variables impacting patient billing within the national healthcare system. As opportunities for predictive modeling expand, their success depends directly on robust data collection, processing, and analysis. Adopting a stringent electronic health record system offers the potential to fully harness machine learning's benefits. Failure to do so risks exacerbating healthcare costs and economic strain amid increasing life expectancy.

APPENDIX

Link to the programming codes for this study on GitHub: https://github.com/bradickristina/HealthInsurance-Cost-Prediction

REFERENCES

[1] Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

[2] Healthdirect Australia (2024, May 30). *The public and private hospital systems* [Text/html], https://www.healthdirect.gov.au/understanding-the-public-and-private-hospital-systems

[3] Hillestad, R., Bigelow, J., Bower, A., Girosi, F., Meili, R., Scoville, R., & Taylor, R. (2005). Can electronic medical record systems transform health care? Potential health benefits, savings, and costs. *Health Affairs (Project Hope)*, *24*(5), 1103–1117. https://doi.org/10.1377/hlthaff.24.5.1103

[4] Hüffmeier, J., Lundman, J., & Elern, F. (2020). Trim and ballast optimization for a tanker based on machine learning. *Ecoprodigi, Intereg Baltic Sea region.*

[5] Kshirsagar, R., Hsu, L.-Y., Greenberg, C. H., McClelland, M., Mohan, A., Shende, W., Tilmans, N. P., Guo, M., Chheda, A., Trotter, M., Ray, S., & Alvarado, M. (2021). Accurate and Interpretable Machine Learning for Transparent Pricing of Health Insurance Plans.

Proceedings of the AAAI Conference on Artificial Intelligence, *35*(17), Article 17. https://doi.org/10.1609/aaai.v35i17.17776

[6] Maisog, J. M., Li, W., Xu, Y., Hurley, B., Shah, H., Lemberg, R., Borden, T., Bandeian, S., Schline, M., Cross, R., Spiro, A., Michael, R., & Gutfraind, A. (2019, December 30). *Using massive health insurance claims data to predict very high-cost claimants: A machine learning approach*. arXiv.Org. https://arxiv.org/abs/1912.13032v1

[7] Orji, U., & Ukwandu, E. (2024). Machine learning for an explainable cost prediction of medical insurance. *Machine Learning with Applications*, *15*, 100516. https://doi.org/10.1016/j.mlwa.2023.100516

[8] Panay, B., Baloian, N., Pino, J. A., Peñafiel, S., Sanson, H., & Bersano, N. (2019). Predicting Health Care Costs Using Evidence Regression. *Proceedings*, *31*(1), Article 1. https://doi.org/10.3390/proceedings2019031074

[9] Panda, S., Purkayastha, B., Das, D., Chakraborty, M., & Kumar Biswas, S. (2022). *Health Insurance Cost Prediction Using Regression Models*. https://ieeexplore.ieee.org/document/9850653

[10] Ray, S. (2019). A Quick Review of Machine Learning Algorithms. *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, 35–39. https://doi.org/10.1109/COMITCon.2019.8862451

[11] Sahai, R., Al-Ataby, A., Assi, S., Jayabalan, M., Liatsis, P., Loy, C. K., Al-Hamid, A., Al-Sudani, S., Alamran, M., & Kolivand, H. (2023). Insurance Risk Prediction Using Machine Learning. *Data Science and Emerging Technologies*, 419–433. https://doi.org/10.1007/978-981-99-0741-0_30

[12] Sahu, A., Sharma, G., Kaushik, J., Agarwal, K., & Singh, D. (2023). *Health Insurance Cost Prediction by Using Machine Learning*. https://doi.org/10.2139/ssrn.4366801

[13] Singh, A., Thakur, N., & Sharma, A. (2016). A review of supervised machine learning algorithms. *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, 1310–1315. https://ieeexplore.ieee.org/abstract/document/7724478

[14] ul Hassan, Ch. A., Iqbal, J., Hussain, S., AlSalman, H., Mosleh, M. A. A., & Sajid Ullah, S. (2021). A Computational Intelligence Approach for Predicting Medical Insurance Cost. *Mathematical Problems in Engineering*, *2021*(1), 1162553. https://doi.org/10.1155/2021/1162553

[15] Vimont, A., Leleu, H., & Durand-Zaleski, I. (2022). Machine learning versus regression modelling in predicting individual healthcare costs from a representative sample of the nationwide claims database in France. *The European Journal of Health Economics*, *23*(2), 211–223. https://doi.org/10.1007/s10198-021-01363-4