Regression-Based Machine Learning Models to Predict Healthcare Costs

Sulaf Elshaar[#1], Saeid Nagem [#2], Saleh Mustafa[#3]

[#]1-2 Faculty of Information Technology, [#]3 Faculty of Sciences
[#]1-3 University of Benghazi- Benghazi, Libya
[1]sulaf.elshaar@uob.edu.ly
[2]saeed.nagem@gmail.com
[3]saleh.mohamed@uob.edu.ly

Abstract— **Accurately predicting healthcare costs is crucial for insurance companies and governments. Machine learning has become increasingly popular in providing more precise methods for predicting healthcare costs. In our research, we conducted several experiments to develop regression models based on healthcare cost prediction. We processed public data from Kaggle and prepared it to train various regression models including Linear Regression (LR), Random Forest Regressor (RFR), and Gradient Boosting (GB). After several training trials, we evaluated the models' performance using multiple methods such as R-squared, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE), along with learning curves, to compare their efficiency and accuracy. We then selected the most optimal model, which was RFR, optimized using GridSearchCV achieved promising results R-squared score of 87.17% in predicting healthcare costs.**

Keywords— Regression models, healthcare cost prediction, data preprocessing, feature selection, performance evaluation metrics.

I. INTRODUCTION

Medical insurance provides a crucial safety net against the financial burdens of unexpected medical expenses. In the United States alone, medical bills are a leading cause of bankruptcy, with studies estimating that around two-thirds of bankruptcies are linked to medical debt [1]. However, individuals with health insurance are significantly less likely to experience medical bankruptcy. For example, a study found that after implementing the Affordable Care Act (ACA), bankruptcy rates among low-income households decreased by over 50% [1].

Access to medical insurance ensures that individuals receive timely and appropriate healthcare services. According to the Center for Disease Control and Prevention (CDC), adults without health insurance are less likely to have a usual source of healthcare and more likely to forgo necessary medical care due to cost concerns [1]. Moreover, uninsured individuals are less likely to receive preventive services, such as cancer screenings and vaccinations, leading to missed opportunities for early detection and intervention. The benefits of medical insurance extend beyond physical health, encompassing mental and emotional well-being. Studies have shown that individuals who gain access to Medicaid experience lower rates of depression compared to those without coverage [1]. Healthcare is a universal need, not limited to a particular country, but government funding and services vary significantly across nations. In 2020, the United States spent a staggering $4.1 trillion on healthcare, with insurance premiums accounting for 31%. This translates to a significant financial burden on individuals and families, with out-of-pocket payments reaching $400 billion [2]. However, in countries where the government does not provide sufficient health care, employees from various sectors often demand that healthcare be provided to them, even if it means deducting part of their salaries. However, insurance companies are reluctant to offer this benefit due to the risk of loss from not being able to accurately estimate the cost of healthcare. Accurate methods to estimate the cost of health care can help insurance companies, and decision-makers predict the expected cost for each individual and provide them with the necessary insurance. Machine learning (ML) is now being used in several areas, including estimating a target value based on a set of inputs. What sets ML apart is its ability to capture nuanced relationships and nonlinear interactions among variables, surpassing the capabilities of traditional statistical methods. Moreover, ML's inherent scalability enables insurers to process large volumes of data efficiently, facilitating comprehensive and robust cost predictions.

ISSN: 2961-6611

By harnessing the predictive ability of ML, Insurance companies can forecast and predict the potential expenses associated with providing coverage to individuals or groups. These predictions play a vital role in risk assessment, premium calculation, resource allocation, and decision-making processes within the insurance industry.

A. *Problem Statement and Motivations*

Medical insurance plays a crucial role in safeguarding people's financial well-being and providing them with access to healthcare. However, insurance companies often struggle to accurately estimate future healthcare costs, which is essential for determining fair premiums for policyholders while ensuring profitability for insurers. Failure to achieve accurate cost estimations can lead to financial instability in the insurance market and adversely affect insurers and insured individuals. To improve healthcare cost estimation, decision-makers should utilize ML. This necessitates the development of ML models that can predict individual healthcare costs.

B. *Aims and Objectives*

This project aims to leverage machine learning's power to develop accurate and reliable models for predicting medical insurance costs. By harnessing the capabilities of algorithms like Linear Regression, Random Forest Regression, and Gradient Boosting, we seek to explore the intricate connections between individuals' characteristics and healthcare expenses. Through rigorous data preprocessing, feature selections, and model optimization techniques, we aim to build robust predictive models that can provide valuable insights to both individuals and insurance providers regarding potential healthcare costs. The project's goals are outlined as follows:

- Cleansing and preprocessing the dataset to prepare it for ML development.
- Constructing predictive models to approximate insurance expenses.
- Evaluating the models' performance using various regression metrics and studying their behaviour.

C. Methodology

To predict medical costs using machine learning, the process starts with acquiring and preparing the required data. We obtained the dataset from Kaggle - a well-known source for obtaining datasets. The initial step in the process is data cleaning, which aims to improve the data's quality and reliability. Afterward, data preprocessing techniques are employed to prepare the dataset for the development of ML models. Once the data is prepared, we proceed to train and optimize the models. The next stage is model evaluation and selection, where various performance metrics are used to assess the trained models' performance. Learning curves are analyzed to gain insights into how the model's performance changes with different datasets.

The remaining sections are organized as follows:

- Related Work is concerned with related work by discussing notable studies that adopted ML in healthcare cost predictions.
- Data Preparation gives insights into the process of preparing the dataset for model training.
- Predictive model development Experiments focus on the ML development task. It demonstrates our experiments in developing and optimizing the predictive models. It also explains how we evaluate the performance of the models and select the best model for predicting the medical cost.
- Conclusion and Future Work presents the most important findings and future work.

II. Related Work

Early research in health insurance cost prediction often used traditional regression algorithms due to their interpretability. Bhardwaj and Anand (2020) compared linear regression, decision trees, random forest, and gradient boosting using personal health data. They found that gradient boosting and multiple linear regression provided the best prediction accuracy, and highlighted the trade-off between accuracy and computational efficiency, with gradient boosting offering faster processing times [3]. This emphasizes the need to balance predictive power and computational cost when choosing algorithms for practical use.

Shyamala Devi et al. (2021) further explored the use of linear and ensemble regression models for health insurance cost prediction, focusing on features such as age, gender, region, smoking status, BMI, and the number of children [4]. Their study demonstrated the effectiveness of ensemble methods, particularly Random Forest, in achieving higher prediction accuracy compared to linear regression models. This aligns with the findings of Bhardwaj and Anand (2020) and suggests that ensemble methods, which combine multiple models to improve predictive performance, hold promise for enhancing the accuracy of health insurance cost prediction.

Feature selection and engineering are vital for enhancing the performance of ML models in cost prediction. Ul Hassan et al. (2021) emphasized the significance of techniques like one-hot encoding for categorical variables and standardization for numerical features in predicting medical insurance costs [5]. Their study underscores the importance of careful data pre-processing and feature transformation to improve the models' ability to identify relevant patterns.

Lui (2012) explored the use of data readily available to insurance companies, such as contract conditions and company characteristics, for predicting employer health insurance premiums [6]. This approach highlights the potential of utilizing existing data sources within the insurance industry to develop predictive models without requiring extensive additional data collection. However, it also raises questions about the limitations of such data in capturing individual-level health risks and utilization patterns, potentially limiting the models' ability to provide personalized cost estimates.

Sailaja et al. (2021) explored the potential of hybrid models by combining multiple regression techniques for medical insurance cost prediction [7]. Their research suggests that hybrid approaches can leverage the strengths of different algorithms to achieve more robust and accurate predictions.

Our project aims to develop accurate regression ML models to predict healthcare costs. We understand the significance of feature engineering and selection and intend to use appropriate techniques to identify the most informative predictors from our chosen dataset. Additionally, optimizing the model's performance is a crucial consideration for us.

III. Data preparation

A. *The Original Dataset*

We integrated two public datasets [9][10] that we found on Kaggle, which were collected in 2023, with a total of 4110 records and seven features. The combined dataset contains essential features for predicting insurance costs like age and gender, health-related factors like Body Mass Index (BMI), smoking status, and the number of children, and the target variable of interest: charges, representing the medical cost of an individual.

B. *Dataset Preprocessing*

Preparing the data for the ML task is a crucial responsibility for ML developers [11]. This includes preprocessing and data cleansing. Hence, we cleaned up our dataset as follows:

- Checking for Missing Values.
- Encoding Categorical Variables (Sex, Smoker, and Region) to numeric variables to make it suitable for the regression model.
- Removing 2773 Duplicated Records.
- Data Scaling and Outliers Detection: we employed the Z-score to detect the outliers. Then, we utilize the StandardScaler() library to standardize our data to ensure that the data points have a balanced scale, which is crucial for many ML algorithms.
- Feature Selection: To examine the relationships between the data features and the target variable, which is "charges" in our case, we used the Univariate Feature Selection technique. Univariate Feature Selection is a common method used for selecting features in ML tasks, particularly in classification or regression problems. This approach evaluates each feature individually against the target variable using statistical tests such as correlation for regression tasks [12] When a feature has a weak correlation with the target, it could decrease model interpretability and efficiency. As a result, our dataset after preprocessing consists of 6 features (age, sex, bmi, children, smoker, and charges) and contains 1337 records.

IV. Predictive Models Development Experiments

A. *Experiments Setup and Work Environment*

This project was primarily developed using Python (version 3.9.18), a versatile programming language well-suited for ML tasks. Python's extensive ecosystem of libraries, including NumPy, Pandas, and Scikit-learn, provides powerful tools for data manipulation, preprocessing, modeling, and evaluation. All training experiments were conducted using a high-performance computer equipped with multicore processors (12 cores), ample RAM (16 GB), and powerful GPUs (6GB).

B. *Model Training*

Choosing suitable ML algorithms influences ML model performance. Therefore, we choose Gradient Boosting (GB), Random Forest Regressor (RFR), and Linear Regression (LR) since they are the most well-known in regression prediction and have demonstrated their ability to produce satisfactory results [19].

After selecting the right algorithms for our research, we carried out several ML model training experiments to achieve an accurate result as follows:

1) Model Training with 70:30 Data Split and Default Parameters: Here, we develop predictive models with default parameter values. Default parameters are predefined values provided by the algorithms. It is worth mentioning that this experiment is meant to serve as a baseline assessment of the performance of ML models. It helps us understand the inherent capabilities of the developed models.
2) Model training with K-fold Cross-Validation: K-fold cross-validation divides the data into K folds, where each fold serves as a validation set while the remaining folds are used for training. The average performance is calculated once this procedure is carried out K times. Not only does cross-validation help to validate ML model results, but it also helps in mitigating overfitting and providing a more accurate estimate of how the model will generalize to unseen data [13][14]. 10-fold Cross-Validation, like in our case, is widely used in the literature to evaluate the performance of ML models [15].
3) Model training with hyperparameter tuning: A hyperparameter is a variable that controls the behavior of the learning algorithm, and Hyperparameter tuning is the process of selecting the ideal set of hyperparameters for an ML model. hence, hyperparameter tuning is an essential part of ML model training. This allows us to optimize model performance for optimal results. We employ both Grid Search and Random Search, which are popular techniques used for hyperparameter tuning [16]. Grid Search systematically explores a predefined grid of hyperparameter values, while Random Search samples hyperparameters from predefined distributions.

C. *Model Performance Evaluation*

Another crucial step in ML development is evaluating the performance of ML models. More specifically, evaluating the model's performance on new, unseen data. We evaluate our predictive models' performance using two techniques, Regression evaluation metrics and learning curves:

1) Regression evaluation metrics: we choose four matrices that are often used in evaluating regression model performance: R-squared (to measure the proportion of the variance in the target variable that is predictable from the independent variables)[17], Mean Absolute Error (MAE) (to measure the average absolute difference between the predicted and actual values), MAE value is returned on the same scale as the target value we are predicting [18], Root Mean Squared Error (RMSE) (to estimate the average deviation between the predicted and actual values in the dataset) [19], and Mean Absolute Percentage Error (MAPE) (to measure the average magnitude of error produced by a model or the average deviation from the predictions)[20].

Because we are required to submit a limited number of pages, we have limited ourselves to presenting the results of our experiments as shown in Table I and Fig.1 (with the default parameters and after applying the GridSearch technique).

TABLE II
MODEL PERFORMANCE ON TEST DATA

	with default parameters				with GridSearchCV			
ML model	R2 (%)	MAE	RMSE	MAPE (%)	R2 (%)	MAE	RMSE	MAPE (%)
LR	77.71	4081.01	6175.57	41.55	77.08	4191.91	6261.65	39.86
RFR	84.00	2957.29	5227.81	33.69	87.17	2590.51	4685.30	28.70
GB	84.64	2946.76	5135.61	32.65	87.09	2620.11	4699.56	31.38

The table III shows that all models have relatively high R-squared values, indicating that they fit our medical dataset and are unbiased. If a model is biased, we cannot trust the results. The GB model has the highest R-squared value (84.64%), followed closely by Random Forest (84%) and Linear Regression (77.71%). A lower MAE indicates better accuracy, which means the model's predictions are closer to the actual values on average. The GB and RFR have the lowest MAE (2946.76). This means their predictions are the closest to the actual medical costs on average. The GB model has the lowest MAPE (32.65%), followed by RFR (33.69%) and LR (41.55%). MAPE calculates the average percentage difference between predicted and actual values. Researchers suggest that the acceptable MAPE value for regression is less than 50. Otherwise, the model predictions are far from the actual target (all our models' MAPE values are below 50) [21]. Among these models, GB and RFR models perform better than LR, as they achieved the lowest values for MAE, RMSE, and MAPE, along with the highest R-squared value. The effectiveness of GridSearchCV on our models is presented in Table IV, RFR and GB have the highest R-squared (87.17 and 87.09, respectively). RMSE suggests that both RFR and GB performed well over the whole dataset.

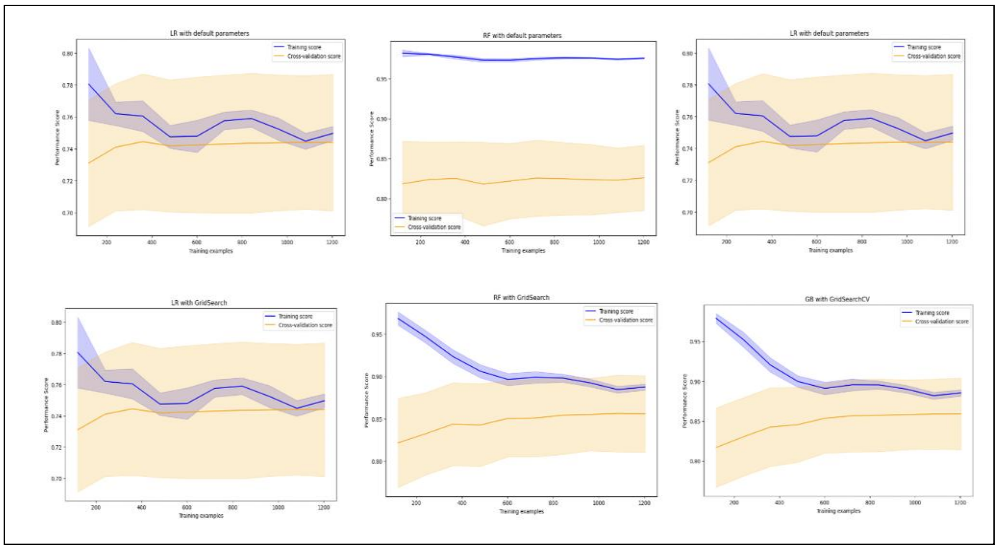

Fig. 1 Learning curves of our ML models (with default parameters and after applying the Gridsearch method)

2) Generating learning curves - to test the model's ability to adapt to new, unseen data. Fig. 1 shows the learning curves generated using training data and validation data. The learning curve for LR explains that the model does not fit very well during the training. RFR fits well only with GridSearch. GB looks well-fitted in all experiments.

V. Results and Discussion

After utilizing four regression metrics (R-squared score, MAE, RMSE, MAPE) and learning curves to evaluate our developed models, we can summarize our findings as follows:

- LR has the lowest performance in terms of R-squared score, MAE, RMSE, MAPE, and learning curve. LR has no hyperparameters that can be affected by hyperparameter tuning, which in our case GridSearch and RandomSearch. This interprets why LR performance was stable in all experiments.
- LR exhibits a slight tendency to overfit as the training score remains relatively high while the cross-validation score dips slightly. This indicates that the model might be memorizing specific patterns in the training data rather than generalizing well to unseen data. For this project, LR is less suitable for predicting the medical cost.
- Both RandomizedSearchCV and GridSearchCV optimized the GB performance by 3%. This means both methods are effective in finding the best hyperparameter settings for GB.
- Regarding RFR, RandomizedSearchCV leads to a slightly lower accuracy (84.41%) compared to GridSearchCV (87.17%). This could be due to the stochastic nature of RandomizedSearchCV, which might not explore the entire hyperparameter space as exhaustively as GridSearchCV.

  RFR and GB exhibit very similar performance in terms of R-squared score, MAE, RMSE, and MAPE. However, the RFR model (developed with GridSearchCV) was closer to the actual medical cost than GB by 3%.
- The optimal predictive model in our case is RFR with GridSearchCV. The regression metrics confirmed that this model is unbiased and produced promising predictions with an accuracy of 87.17%. Moreover, after examining its learning behavior when tested on validation data, we found that it was less prone to overfitting.

VI. Conclusion and Future Work

Accurately predicting insurance costs is essential for the insurance and healthcare sectors. This project focused on predicting medical insurance costs using machine learning techniques. We examined various algorithms, including Linear Regression, Random Forest Regressor, and Gradient Boosting, each with distinct strengths in modeling relationships between individual characteristics and healthcare expenses. Through data preprocessing, feature selection, and model optimization like cross-validation and hyperparameter tuning, we developed effective predictive models. We evaluated model performance using regression metrics such as R-squared, Mean Absolute Error, and Root Mean Squared Error. Our results showed that Random Forest Regressor and Gradient Boosting outperformed Linear Regression, with the Random Forest model achieving a notable R-squared score of 87.17%. While this project highlights the potential of machine learning in medical insurance cost prediction, there's room for improvement, such as incorporating additional data sources and exploring advanced algorithms. Overall, the findings showcase the transformative impact of machine learning in the insurance industry, providing valuable tools for informed decision-making in healthcare finance.

References

[1] American Hospital Association, "Report: The Importance of Health Coverage | AHA."

[2] Organisation for Economic Co-operation and Development, "OECD Health Statistics 2023 - OECD."

[3] N. Bhardwaj, R. A. Delhi, I. D. Akhilesh, and D. Gupta, "*Health Insurance Amount Prediction," International Journal of Engineering Research & Technology*, vol. 9, no. 5, May 2020, doi: 10.17577/IJERTV9IS050700.

[4] M. Shyamala Devi et al., "*Linear and Ensembling Regression Based Health Cost Insurance Prediction Using Machine Learning," Smart Innovation, Systems and Technologies,* vol. 224, pp. 495–503, 2021

[5] C. A. ul Hassan, J. Iqbal, S. Hussain, H. AlSalman, M. A. A. Mosleh, and S. Sajid Ullah, "A Computational Intelligence Approach for Predicting Medical Insurance Cost," Math Probl Eng, vol. 2021, 2021, doi: 10.1155/2021/1162553.

[6] E. Lui, "Employer Health Insurance Premium Prediction," CiteCeerX, 2012.

[7] N. Venkata Sailaja, M. Karakavalasa, M. Katkam, M. Devipriya, M. Sreeja, and D. N. Vasundhara, "HYBRID REGRESSION MODEL FOR MEDICAL INSURANCE COST PREDICTION AND RECOMMENDATION," Proceedings - 2021 IEEE International Conference on Intelligent Systems, Smart and Green Technologies, ICISSGT 2021,

[8] R. HENCKAERTS, "Insurance pricing in the era of machine learning and telematics technology," 2021.

[9] N. Ashraf, "Health Insurance Cost Prediction."

[10] H. KUMAR, "Medical Insurance Price Prediction: Predicting Medical Insurance Prices: A Data-Driven Approach to Forecasting Heal."

[11] Elshaar, S., Sadaoui, S. " Cost-Sensitive Semi-supervised Classification for Fraud Applications." In: Rocha, A.P., Steels, L., van den Herik, J. (eds) Agents and Artificial Intelligence. ICAART 2020. Lecture Notes in Computer Science(), vol 12613. Springer, Cham, 2021, https://doi.org/10.1007/978-3-030-71158-0_8

[12] P. Schober and L. A. Schwarte, "Correlation coefficients: Appropriate use and interpretation," Anesth Analg, vol. 126, no. 5, pp. 1763–1768, May 2018, doi: 10.1213/ANE.0000000000002864.

[13] M. Kuhn and K. Johnson, "Applied Predictive Modeling," Springer, 2013.

[14] S. Shalev-Shwartz and S. Ben-David, Understanding Machine Learning: From Theory to Algorithms, First. Cambridge University Press, 2014. [Online]. Available: http://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning

[15] T. Gunasegaran and Y. N. Cheah, "Evolutionary cross validation," ICIT 2017 - 8th International Conference on Information Technology, Proceedings, pp. 89–95, Oct. 2017, doi: 10.1109/ICITECH.2017.8079960.

[16] J. Bergstra and Y. Bengio, "Random Search for Hyper-Parameter Optimization Yoshua Bengio," Journal of Machine Learning Research, vol. 13, pp. 281–305, 2012, [Online].

[17] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE, and RMSE in regression analysis evaluation," *PeerJ Comput Sci*, vol. 7, pp. 1–24, Jul. 2021

[18] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature," *Geosci Model Dev*, vol. 7, no. 3, pp. 1247–1250, Jun. 2014, doi: 10.5194/GMD-7-1247-2014.

[19] S. K. K. Babu, "Mathematical Modelling of RMSE Approach on Agricultural Financial Data Sets," *Int J Pure Appl Biosci*, vol. 5, no. 6, pp. 942–947, Dec. 2017, doi: 10.18782/2320-7051.5802.

[20] A. de Myttenaere, B. Golden, B. Le Grand, and F. Rossi, "Mean Absolute Percentage Error for regression models," *Neurocomputing*, vol. 192, pp. 38–48, Jun. 2016, doi: 10.1016/J.NEUCOM.2015.12.114.

[21] J. J. Montaño Moreno, A. Palmer Pol, A. Sesé Abad, and B. Cajal Blasco, "Using the R-MAPE index as a resistant measure of forecast accuracy," *Psicothema*, vol. 25, no. 4, pp. 500–506, 2013, doi: 10.7334/PSICOTHEMA2013.23.

ISSN: 2961-6611