Medical Insurance Price Prediction Using Machine Learning

Bandi, Om Preetham
College of Engineering & Computer Science
Texas A & M University - Corpus Christi
Corpus Christi, Texas USA

Polsani, Akhil
College of Engineering & Computer Science
Texas A & M University - Corpus Christi
Corpus Christi, Texas USA

Ullendula, Thriveen
College of Engineering & Computer Science
Texas A & M University - Corpus Christi
Corpus Christi, Texas USA

Virigineni, Sravya Sri
College of Engineering & Computer Science
Texas A & M University - Corpus Christi
Corpus Christi, Texas USA

December 7, 2023

Abstract

The escalating costs of healthcare insurance present a significant financial challenge to individuals and institutions alike, rendering the ability to accurately predict these expenses critical for budgetary and policy considerations. Motivated by this exigency, our paper explores the application of machine learning techniques to predict individual insurance charges based on demographic and health-related features. We utilized an extensive dataset, implementing an XGBoost regression model known for its high efficiency and predictive accuracy. The model was trained and validated on features including age, BMI, smoking status, and number of children, which were determined to be influential through an exploratory data analysis. Our methodology included rigorous preprocessing, feature selection based on importance scores, and hyperparameter tuning to optimize the model's performance. The results were promising, showcasing the model's capability to predict insurance costs with a high degree of precision, evidenced by a training accuracy of 0.871, a test accuracy of 0.904, and a cross-validation score of 0.860. The findings from this study not only contribute to the existing body of knowledge but also have practical implications for stakeholders in the healthcare insurance domain.. . .

1 Introduction

Background The realm of personal finance is ever evolving, and among its many facets, the cost of healthcare insurance remains a significant concern for both individuals and policymakers. Accurate prediction of insurance premiums stands as a pillar for financial planning and healthcare affordability. Considering this, our study targets the development of a predictive model that can estimate insurance costs based on personal health data and demographics.

Motivation The landscape of healthcare insurance is particularly complex due to the multitude of influencing factors ranging from individual health indicators to broader socioeconomic conditions. The exploration of this domain is not just academically stimulating but also socially pertinent, given the rising trajectory of healthcare expenses. Thus, our motivation is anchored in creating a tool that brings clarity and predictability to insurance cost estimation.

Problem The impetus for addressing this problem is twofold: to aid individuals in anticipating healthcare expenses and to assist insurance companies in establishing fair and transparent pricing structures.

Algorithm For this investigation, we have utilized a machine learning approach, capitalizing on the robust capabilities of ensemble learning via an XGBoost regression model. The inputs to our algorithm are structured data points, specifically: age, body mass index (BMI), number of children, smoking status, and geographical region of the policyholder. These inputs are reflective of both the policyholder's personal attributes and external factors likely to impact insurance charges. The output of our model is a single value: the predicted annual insurance premium for the policyholder. This predictive output is expected to serve as a cornerstone for financial decision-making processes for prospective insurance buyers and provide a benchmark for insurers in rate-setting exercises.

Outline The remainder of this article is organized as follows. Section 2 gives account of previous work, followed by the Section 3, which deep dives into the dataset and its features. Section 4 explains about the algorithms and final model. Our new and exciting results are described in Section 5. Finally, Section 6 gives the conclusions and future implementations.

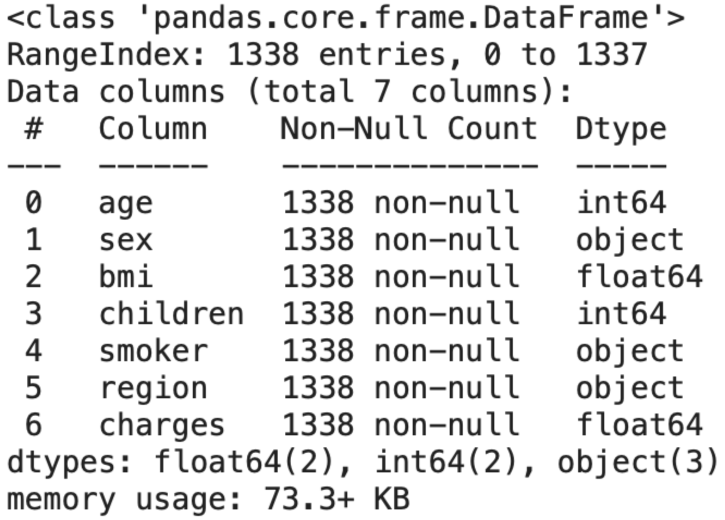
2 Related work

The study [1], demonstrates the efficacy of machine learning regression techniques, particularly Gradient Boosting and AdaBoost, in predicting medical insurance costs based on patient data, outperforming traditional models like Lasso and Elastic Net Regression. [9] employs linear regression, naive Bayes, and random forest algorithms to assess healthcare costs, particularly emphasizing the role of BMI, with linear regression emerging as the most accurate. The thesis [2] compares Multiple Linear Regression, Decision Tree Regression, and Gradient Boosting Decision Tree Regression, finding the latter two, especially Gradient Boosting, to be more effective in predicting insurance costs. [3], the application of machine learning in clinical settings is explored, discussing its advantages over traditional methods, the challenges involved, and potential future applications in healthcare [7]. From Computers, Materials & Continua, introduces new ensemble methods for insurance cost prediction, with a focus on boosting [10] and stacking ensembles which show better accuracy and lower error rates compared to bagging ensembles. We used numpy[4], pandas[6], seaborn[11], matplotlib[5], warnings[8] for this project.

3 Dataset and Features

The dataset employed in this study provides a comprehensive overview of individual health-related attributes and their associated medical insurance costs. The dataset, sourced from a reputable healthcare analytics repository Kaggale, contains several key features. These include age, BMI, and the number of children, all of which are continuous variables that have been discretized appropriately for use in our model. The feature 'age' was treated as a continuous integer, 'BMI' was processed as a floating-point number representing the body mass index value, and 'children' was considered an integer count of the dependents. It comprises 1,338 instances, which have been divided into a training set of 1,071 (80%) and a validation/test set of 267 (20%). This partition ensures a robust training process while reserving a substantial portion of data for the critical evaluation of our model's generalization capabilities.

Data Loading and Quality Assessment Upon loading the data using pandas, an initial inspection revealed a well-structured dataset with no missing values, as confirmed by the isnull().sum() method. The df.info() as shown in Figure 1 command provided an overview of the data types and non-null count, indicating a mix of numerical and categorical variables.

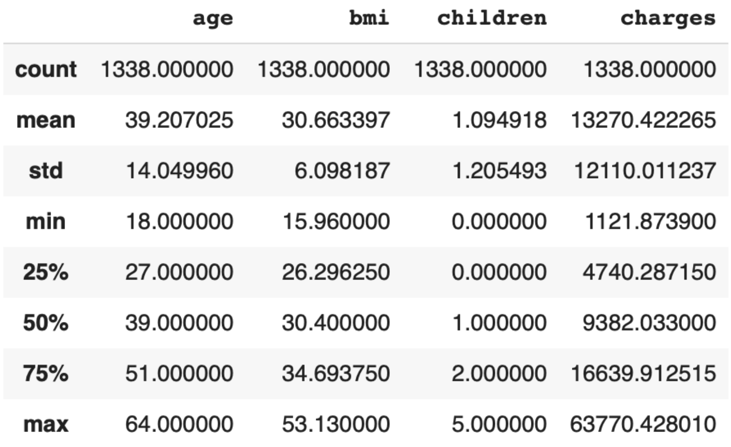


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1338 non-null   int64
 1   sex       1338 non-null   object
 2   bmi       1338 non-null   float64
 3   children  1338 non-null   int64
 4   smoker    1338 non-null   object
 5   region    1338 non-null   object
 6   charges   1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB
```

Figure 1: Details about the columns of the dataset.

Statistical Overview We performed a statistical summary using df.describe() as shown in Figure 2, giving us insight into the central tendencies and variations within numerical features like age, BMI, and insurance charges. Notably, the presence of outliers was identified, particularly within the 'bmi' feature.

Outlier Management To address outliers and ensure they did not skew our model, we employed the Interquartile Range (IQR) method as shown in Figure 3, capping extreme values within a defined threshold. This treatment was visualized through boxplots, confirming the effective mitigation of outlier influence.



	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000
75%	51.000000	34.693750	2.000000	16639.912515
max	64.000000	53.130000	5.000000	63770.428010

Figure 2: Descriptive statistical measures of the dataset.

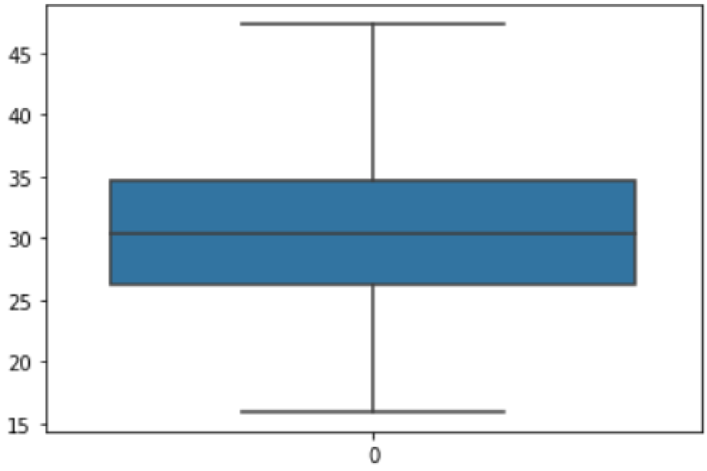


Figure 3: Boxplot of bmi after outlier treatment.

Categorical Variable Analysis Our categorical variables—’sex’, ’smoker’, and ’region’—were visualized using pie charts, exposing an imbalance in smoker status distribution, a critical variable likely to impact insurance costs significantly.

Correlation Analysis We conducted a correlation analysis to explore the relationships between features. The resulting correlation matrix aided in identifying multicollinearity and the degree of association between variables.

Feature Encoding Prior to model training, the dataset underwent a series of preprocessing steps. All categorical variables, including sex, smoker status, and region, were transformed into numerical values to facilitate their use in our algorithms. Specifically, ’sex’ was encoded as 0 for male and 1 for female, ’smoker’ was encoded as 1 for a smoker and 0 for a non-smoker, and ’region’ was mapped to a numerical scale ranging from 0 to 3 representing the northwest, northeast, southeast, and southwest regions respectively.

Data Split The dataset was partitioned into a training set and a validation/test set, adhering to an 80-20 split ratio. This ensured a sufficient amount of data for model training while reserving an adequate portion for model evaluation.

In terms of normalization, we implemented the Z-score standardization to all continuous features, ensuring that each feature contributed proportionately to the final prediction. This step was critical in mitigating the impact of differing scales and variances across the various features.

4 Methods

In this study, our primary focus was the application of the XGBoost (eXtreme Gradient Boosting) algorithm for predicting medical insurance costs. XGBoost is a decision-tree-based ensemble machine learning algorithm that uses a gradient boosting framework. Renowned for its performance and speed, it has been a popular choice in data science competitions and practical applications. Along with XGBoost we also worked and tested with Linear Regression, SVR, Gradient Boosting, Random Forest.

4.1 Linear Regression

Functioning: Models the relationship between a dependent variable and one or more independent variables using a linear equation.
Mathematical Formulation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon \tag{1}$$

where y is the dependent variable, x_i are independent variables, β_i are coefficients, and ϵ is the error term.
Loss Function: Mean Squared Error (MSE),

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{2}$$

4.2 Support Vector Machine (SVM) for Regression (SVR)

Functioning: Applies SVM principles to regression, aiming to find a function that minimizes deviation from observed outputs within a margin.
Mathematical Formulation: SVR solves the following optimization problem:

$$\min_{w,b,\zeta,\zeta^*} \frac{1}{2} w^T w + C \sum_{i=1}^{n} (\zeta_i + \zeta_i^*) \tag{3}$$

subject to constraints for error margin and regularization.
Loss Function: Epsilon-Insensitive Loss.

4.3 Random Forest

Functioning: An ensemble method using multiple decision trees to perform regression.
Mathematical Formulation:

$$f(x) = \frac{1}{B} \sum_{b=1}^{B} f_b(x) \tag{4}$$

where B is the number of trees and $f_b(x)$ is the prediction of the b-th decision tree.
Loss Function: Typically Mean Squared Error (MSE).

4.4 Gradient Boosting

Functioning: Constructs a predictive model in a stage-wise manner, correcting previous errors with new trees.
Mathematical Formulation:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \tag{5}$$

where $F_{m-1}(x)$ is the accumulated model until iteration $m - 1$, $h_m(x)$ is the new tree, and γ_m is the learning rate.
Loss Function: A differentiable loss function, often MSE or Mean Absolute Error (MAE).

4.5 XGBoost Algorithm

XGBoost operates by constructing a multitude of decision trees in a sequential manner. Each new tree corrects the errors made by the previously built trees. The model's output is a combination of the predictions from all the trees. The prediction model in XGBoost is formalized as:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \tag{6}$$

where $\hat{y}_i$ is the predicted output for the i^{th} instance, f_k represents an individual decision tree, and x_i denotes the feature vector for the i^{th} instance.

Objective Function XGBoost's objective function is a combination of a loss function and a regularization term. The objective function is given by:

$$\text{Obj} = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{7}$$

Here, l is the loss function, and Ω penalizes the complexity of the model.

Loss Function For regression tasks, the mean squared error (MSE) is used as the loss function, defined as:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{8}$$

where n is the number of instances in the dataset.

4.6 Comparing All Models

Table 1: Model Performance Comparison

Model	**Train Accuracy**	**Test Accuracy**	**CV Score**
LinearRegression	0.729	0.806	0.747
SupportVectorMachine	-0.105	-0.134	0.103
RandomForest	0.974	0.882	0.836
GradientBoost	0.868	0.901	0.860
XGBoost	0.870	0.904	0.860

Here, we can see a comparison of models like Linear Regression, SVR, Random Forest, Gradient Boosting, and XGBoost. We evaluated them based on their accuracy and R² scores. After careful analysis, XGBoost stood out as the top performer, demonstrating the highest accuracy and R² score among all the models tested.

5 Experiments/Results/Discussion

5.1 Experiments

Our experimental approach was methodical, focusing on both quantitative and qualitative aspects of the model's performance. We meticulously tuned hyperparameters and evaluated the model using relevant metrics.

Now we need to identify the important features for predicting of charges.

Table 2 illustrates the importance scores assigned to each feature.

Notice how certain factors like smoker status and age stand out. One of the interesting findings was the high importance of smoker status, which significantly influences insurance premiums.

	Importance
age	0.050547
sex	0.002721
bmi	0.092197
children	0.013500
smoker	0.834279
region	0.006756

Table 2: Feature Importance

Through feature importance analysis, we identified and excluded less impactful features like 'sex' and 'region', streamlining the model for efficiency and focus. As shown in Table 3.

Hyperparameter Tuning

	Importance
age	0.050547
bmi	0.092197
children	0.013500
smoker	0.834279

Table 3: Feature Importance without Sex and Region

XGBoost Parameters The key hyperparameters adjusted for the XGBoost model included:

- **n_estimators** Set to 100, balancing model complexity and training time.
- **max_depth** Capped at 3 to prevent overfitting while allowing the model to capture underlying patterns.
- **learning_rate** Set at 0.1, a conservative rate to ensure gradual learning.
- **gamma** Kept at a minimal value to prioritize minimizing the loss function over making the model too simplistic.

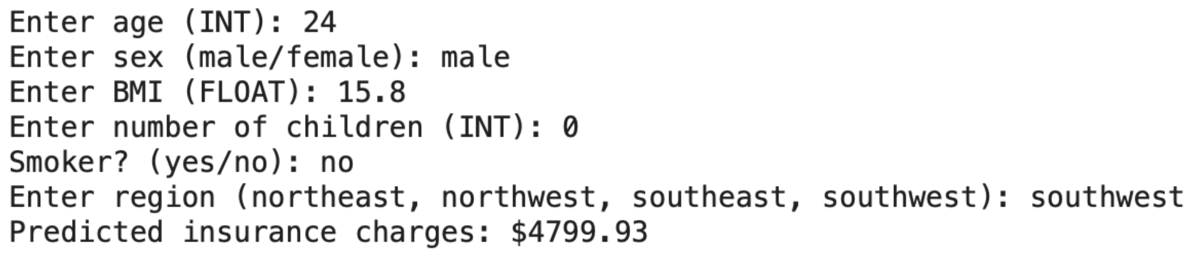
Rationale These parameters were chosen through a grid search approach, where multiple combinations were tested and evaluated against the cross-validation set.

5.2 Results

- **Primary Metrics:**
  - The primary metrics for evaluating our regression model were accuracy and R² Score.
  - Equations:
    $$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$
    where $\bar{y}$ is the mean of observed data.
- **Results**
  - The XGBoost model achieved an Train accuracy of 0.870691899927822, Test accuracy of 0.904151903449132 and an R² score of 0.90.
  - Then we tested our model and we almost got perfect predicted insurance.

5.3 Discussion

- The model demonstrated high accuracy, as indicated by the R^2 score.
- **Qualitative Analysis:** The feature importance plot revealed that 'smoker' status was the most influential predictor, aligning with known trends in insurance pricing.
- **Model Evaluation:** Residual plots showed a random dispersion of residuals, indicating that the model did not systematically under or over-predict across the range of data.



```
Enter age (INT): 24
Enter sex (male/female): male
Enter BMI (FLOAT): 15.8
Enter number of children (INT): 0
Smoker? (yes/no): no
Enter region (northeast, northwest, southeast, southwest): southwest
Predicted insurance charges: $4799.93
```

Figure 4: Predicting Medical Insurance for person.

Algorithm Performance and Overfitting

- Compared to other algorithms like Linear Regression and SVM, XGBoost showed superior performance in handling non-linearities and interactions between features.
- To check for overfitting, we monitored the performance gap between training and validation sets. The model's consistent performance across both sets suggested minimal overfitting.
- Regularization parameters in XGBoost, like `gamma` and `max_depth`, played a crucial role in preventing overfitting.

6 Conclusion/Future Work

6.1 Conclusion

Our study embarked on a journey to develop a machine learning model capable of accurately predicting medical insurance costs. The project involved evaluating various algorithms, with XGBoost emerging as the standout performer. This conclusion was reached after rigorous testing and comparison, which involved Linear Regression, SVM, Random Forest, Gradient Boosting, and ultimately, XGBoost. The key to XGBoost's success lay in its ability to handle complex, non-linear relationships within the data, a feature that was not as effectively managed by simpler models like Linear Regression or SVM. The ensemble approach of Random Forest and Gradient Boosting provided substantial improvements over single-estimate models, but XGBoost's sophisticated boosting and regularization techniques offered the best balance of accuracy and overfitting control.

Key Points

- XGBoost's superior performance was evidenced by its high R² score.
- Feature importance analysis revealed critical predictors like smoking status, which aligned with intuitive expectations about insurance cost drivers.
- The model demonstrated robustness across training and validation datasets, indicating its generalizability.

6.2 Future Work

Incorporating Additional Data Expanding the dataset to include more variables, such as medical history or lifestyle factors, could provide deeper insights and improve the model's predictive power. Exploring Advanced Models: Leveraging deep learning algorithms, especially neural networks, might uncover more intricate patterns within the data.

Refining Hyperparameter Tuning With more computational resources, a more exhaustive grid search or advanced optimization techniques like Bayesian Optimization could fine-tune the model further.

Real-Time Predictions Developing an application for real-time insurance cost predictions, integrating the model into an interactive platform.

Cross-Domain Applications Adapting the model for use in related domains, such as life or vehicle insurance, to test its versatility and effectiveness across different types of insurance cost predictions.

7 Contributions

- Bandi, Om Preetham:
  - Om Preetham's proficiency in coding was instrumental in data preprocessing and model development. He took charge of writing the Python scripts necessary for cleaning the dataset and implementing the machine learning algorithms.
  - His contributions were critical in the initial stages of our project pipeline. He produced the graphs and charts that helped the team understand the data, and his visualizations played a significant part in the presentation and report, making complex data more accessible.
- Polsani, Akhil:
  - Akhil brought a strong analytical perspective to the team, leading the statistical analysis part of the project.
  - He worked closely with Om Preetham to ensure the data was correctly processed and played a key role in interpreting the results of our models, providing the team with actionable insights.
- Ullendula, Thriveen:
  - With a keen eye for detail, Thriveen focused on the fine-tuning of our machine learning models.
  - He experimented with various hyperparameters to optimize model performance, ensuring that our predictions were both accurate and reliable.

- Virigineni, Sravya Sri:
    - Sravya's contributions to the coding aspect of the project were paramount.
    - She collaborated with Om Preetham and Thriveen on implementing the machine learning algorithms and provided essential support in debugging and refining the codebase.

References

[1] Haitham M. Alzoubi, Nizar Sahawneh, Ahmad Qasim AlHamad, Umar Malik, Ameer Majid, and Ayesha Atta. Analysis of cost prediction in medical insurance using modern regression models. In *2022 International Conference on Cyber Resilience (ICCR)*, pages 1–10, 2022.

[2] Nidhi Bhardwaj and Rishabh Anand. Health insurance amount prediction. *Int. J. Eng. Res*, 9:1008–1011, 2020.

[3] Alison Callahan and Nigam H Shah. Machine learning in healthcare. In *Key advances in clinical informatics*, pages 279–291. Elsevier, 2017.

[4] C.R. Harris, K.J. Millman, S.J. van der Walt, et al. Array programming with numpy. *Nature*, 585:357–362, 2020.

[5] John D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[6] Wes McKinney. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, pages 51–56, 2010.

[7] Thais Carreira Pfutzenreuter and EP Lima. Machine learning in healthcare management for medical insurance cost prediction. 2021.

[8] Python Software Foundation. Warnings control. https://docs.python.org/3/library/warnings.html.

[9] Rahul Sahai, Ali Al-Ataby, Sulaf Assi, Manoj Jayabalan, Panagiotis Liatsis, Chong Kim Loy, Abdullah Al-Hamid, Sahar Al-Sudani, Maitham Alamran, and Hoshang Kolivand. Insurance risk prediction using machine learning. In *The International Conference on Data Science and Emerging Technologies*, pages 419–433. Springer, 2022.

[10] Ahmed I Taloba, Abd El-Aziz, M Rasha, Huda M Alshanbari, Abdal-Aziz H El-Bagoury, et al. Estimation and prediction of hospitalization and medical care costs using regression in machine learning. *Journal of Healthcare Engineering*, 2022, 2022.

[11] Michael Waskom. Seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021.