Prediction, Comparison and Visualization of Premium Pricing Using Machine Learning Methods

Farzan Tashfeen[1], Mohid Hussain[2] Haitham Ahmer[3] , …….
[1] NED University of Engineering and Technology
[2] NED University of Engineering and Technology
[3] Industrial Institute of Electronic Engineering

Abstract

This study delves into the effectiveness of machine learning in forecasting and comparing health insurance premiums. The all-embracing goal is to improve the accuracy of premium pricing and minimize the manual labor associated with the traditional methods and present data trends in a comprehensible manner. Utilizing an open-source dataset of US medical premiums, the authors have developed a meticulous machine learning model in Python programming language. This model took into account factors like age, bmi, smoking habits, and geographical location. The outcomes of this research accentuate the transformative potential of machine learning in offering precise pricing that stands to the advantage of both insurer and policyholder. Remarkably, XGBoost excelled other models with an RMSE value of 0.349 and an adjusted R^2 of 90.1% and this was closely trailed by the author's Artificial Neural Network model which achieved an accuracy of 93.07%. The outcomes of the research findings affirm the central hypothesis that machine learning offers a robust and efficient methodology for accurate premium pricing to the advantage of both the insurer and policyholder.

Keywords: Premium Pricing, Prediction, Health Insurance, Model, Machine Learning, Artificial Intelligence

1. Introduction

Health insurance shields individuals from the financial burdens of medical bills. Upon payment of a certain premium, an individual can secure coverage under a health insurance policy [1−7]. The computation of this premium is influenced by a multitude of factors, which leads to a variability in the cost of health insurance policies. For instance, age significantly impacts the cost of a health insurance plan [8−13]. Younger individuals naturally have fewer health issues than older ones and normally require less extensive medical treatment. As such, the cost of treating an older person is higher and consequently they are charged a higher premium. Multiple factors influence the insurance premium of a health insurance policy that leads to a unique premium amount for each individual. In recent years, the healthcare industry has witnessed a surge in the application of artificial intelligence (AI), performing medical-related tasks at a much faster rate [14−16]. The ability of AI to gather, process and provide appropriate results from data expedites disease detection and minimizes errors, thereby substantially shortening the diagnosis-treatment-recovery cycle. For instance, healthcare professionals and organizations utilize chatbots to gather basic patient information before a consultation, improving efficiency for both the doctor and the patient

[17−21]. In fact, AI and machine learning (ML) have made significant advances in the health insurance sector. AI facilitates faster claim settlements by detecting fraudulent claims, enabling insurers to process valid claims more efficiently [22]. Additionally, the potential of AI to recommend healthier habits and behaviors to clients offers cost-effectiveness to insurers thereby reducing avoidable healthcare expenditures. The traditionally lengthy and time-consuming process of underwriting has been expedited by AI-based predictive analysis. Research in the field of AI with regard to healthcare can enhance efficiency, improve treatment efficacy and reduce burnout among medical professionals. The AI and ML technologies have the competency to expedite insurance claim processing, reduce the burden of electronic health records and provide individual health insurance plans and, as such, transform the healthcare experience in developing countries [23−24]. Even though there is considerable advancement in AI and ML in the domain of healthcare, the scene of healthcare even then is confronted with a unique set of challenges in terms of the complex data evaluation and the habitually high patient costs. This study leveraged ML models in Python programming language intending to predict, compare and visualize the health insurance premiums. The models will help reduce manual pricing labor, improve the accuracy and visualize data for decision-making to the benefit of both the agencies and customers.

2. Literature Review

The landscape of premium pricing prediction using machine learning (ML) methods has been explored in a number of research studies. A literature appraisal of the research work done in this realm reveals that the scope of utilizing ML algorithms for premium pricing prediction is vast and multi-faceted. The research work done by Keshav Kaushik, Akashdeep Bhardwaj, Ashutosh Dhar and Dwivedi Rajani Singh [25] trained and evaluated an artificial intelligence network-based regression model to predict health insurance premiums. The authors predicted the health insurance cost incurred by individuals on the basis of a host of parameters like age, gender, BMI (body-mass-index), number of children, smoking habits, and geographical location with a consequential accuracy of 92.72%. The research team of Ch. Anwar ul Hassan, Jawaid Iqbal, Saddam Hussein, Hussain Al Salman, Mogeeb A. A. Mosleh and Syed Sajid Ullah [26] exploited a variety of machine learning algorithms to predict medical insurance costs. They calculated the skewness and kurtosis for each attribute so as to check the outliers. Their research encompassed data preprocessing, feature engineering, regression and evaluation. In their findings, the Stochastic Gradient Boosting model stood out prominent and claimed an accuracy of 86%. Takeshima T, Keino S, Aoki R, Matsui T and Iwasaki K [27], developed a prediction model using health insurance claims data. They applied the least absolute shrinkage and selection operator (LASSO) regression model and used excess logarithmic medical cost as explained valuables and diseases as explanatory valuables in order to avoid overfitting. In their research work, Sen Hu, Adrian O'Hagan, James Sweeney, and Mohammadhossein Ghahramani [28] analyzed lapse behavior in life insurance policies. The incorporation of spatially linked demographic census data provided insights for insurers to assess and prevent lapsing risks, enhancing the calculation and preparation

of capital reserves. In the same vein, Angela D. Kafuria [29], developed a predictive model for determining health insurance charges using machine learning algorithms. The model was trained on attributes like age, sex, bmi, number of children, and region. They highlighted in their research that Gradient Boosting and Random Forest Regression are the most effective predictive models. Nonetheless, the work of Y. Angeline Christobel and Suresh Subramanian [30] compared four regression analysis algorithms to predict insurance namely Linear Regression, Ridge Regression, Lasso Regression and Polynomial Regression. They concluded in their research findings that Among the four models the Polynomial Regression emerged as the most effective model with an accuracy of 88%.

The literature survey divulges the fact that ML models hold significant promise in predicting health insurance premiums that can revolutionize the insurance industry in terms of predicting medical costs and analyzing the customer behavior. Nevertheless, predicting health insurance premiums in the healthcare business by incorporating ML models and algorithms is still a theme that has to be investigated, understood and improved. Accordingly, in this research endeavor the authors are going to examine the ML models in the programming language of Python for an advanced prediction, comparison and visualization of the health insurance premiums. It is expected that the models will help reduce manual pricing labor, improve the accuracy and picture the data for decision-making to the advantage of both the agencies and customers.

3. Research Methodology

The research methodology for this study is made up of of several steps. The authors started with data import and analysis, then did feature engineering and data visualization and finally developed the model and tested the model for predictions.

3.1. Data Import and Preprocess

In this study we utilized the US Health Insurance Dataset which has been employed previously in several research studies among which the prominent research studies are undertaken by researchers like Kaushik et al [25], Hassan et al [26] and Takeshima et al [27]. This dataset is widely recognized for its comprehensiveness and the key attributes like age, sex, BMI, the number of children, smoking habits, and regional details are tested and judged against the insurance charges. Our decision to use this dataset is based on its proven reliability and relevance to our research objectives. In order to remove duplicates the dataset was preprocessed and cleaned.

3.2. Feature Engineering

We transformed the categorical columns into binary formats for model compatibility. Precisely, the sex column was converted to binary, where 0 denotes female and 1 indicates male. Likewise, the smoker column was binarized with 0 for no and 1 for yes. The region column was feature engineered where the categorical entries were transformed using dummy variables and were then

appended to the main dataset. This way, the original 'region' column was removed to avoid redundancy.

3.3. Exploratory Analysis

A comprehensive histogram was generated to visualize the distribution of all the attributes in the dataset including age, sex, bmi, children, smoker, charges and regional information (i.e. Northwest, Southeast, Southwest). This initial graphical examination helped in understanding the data distribution for each attribute.

A multi-faceted visualization was designed to explore the distribution of three key variables, i.e. age, bmi and charges. This was achieved using box plots, violin plots, and histograms for each variable. Then a bivariate map was generated to display the correlation coefficients between all the numerical attributes. To further gain a deeper understanding of the relationships between different features in the dataset, a Pair Grid plot was utilized. This Grid provides a matrix of plots where each feature is plotted against every other feature.

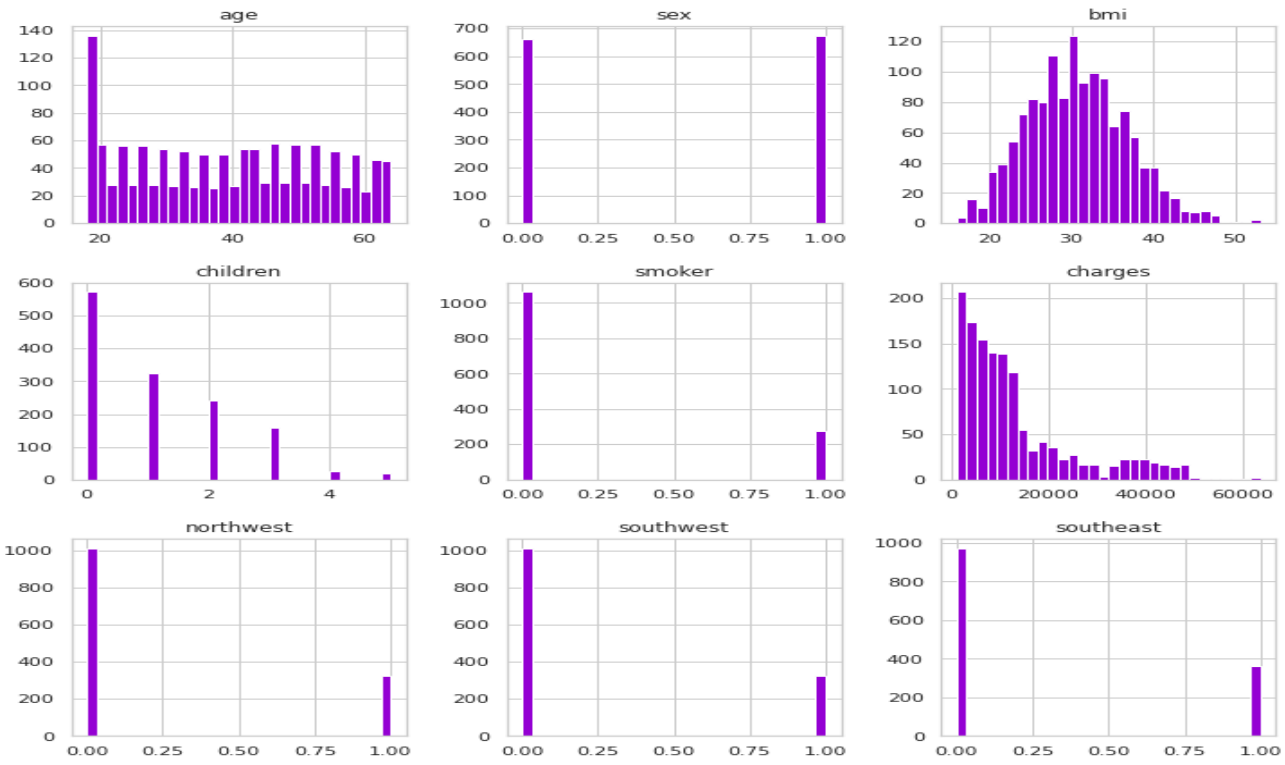

Figure 1. Histograms pictures for the distribution of Attributes

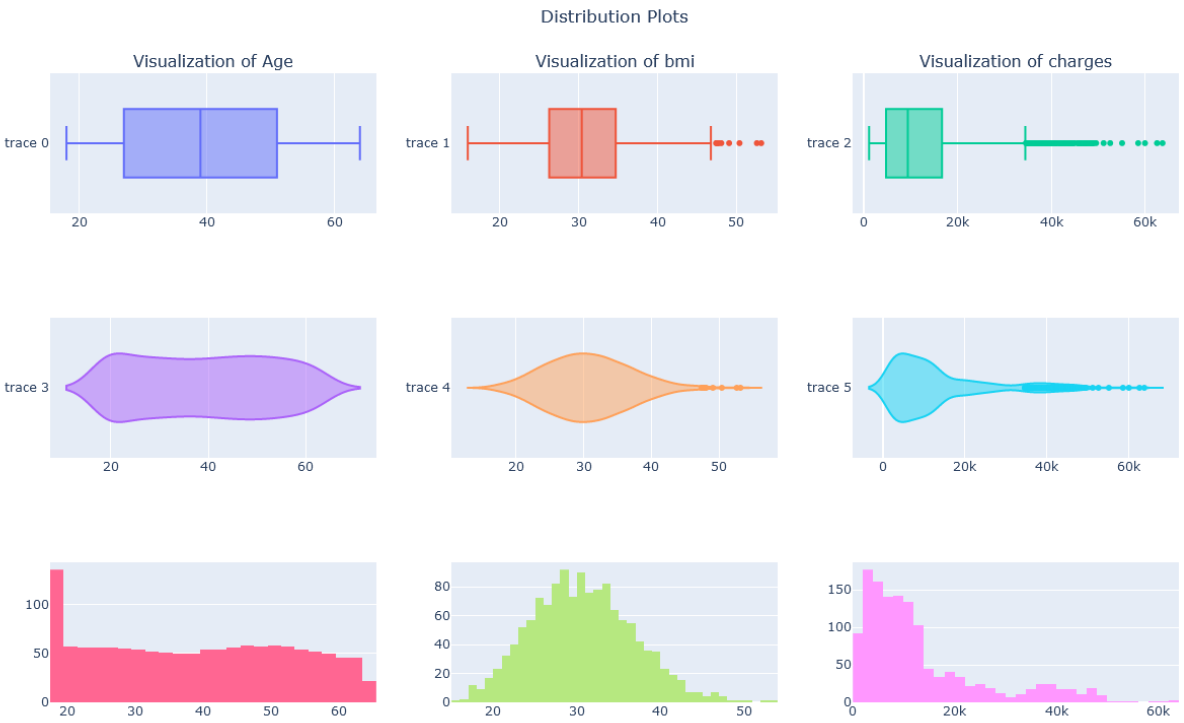

Figure 2. Distribution Plots

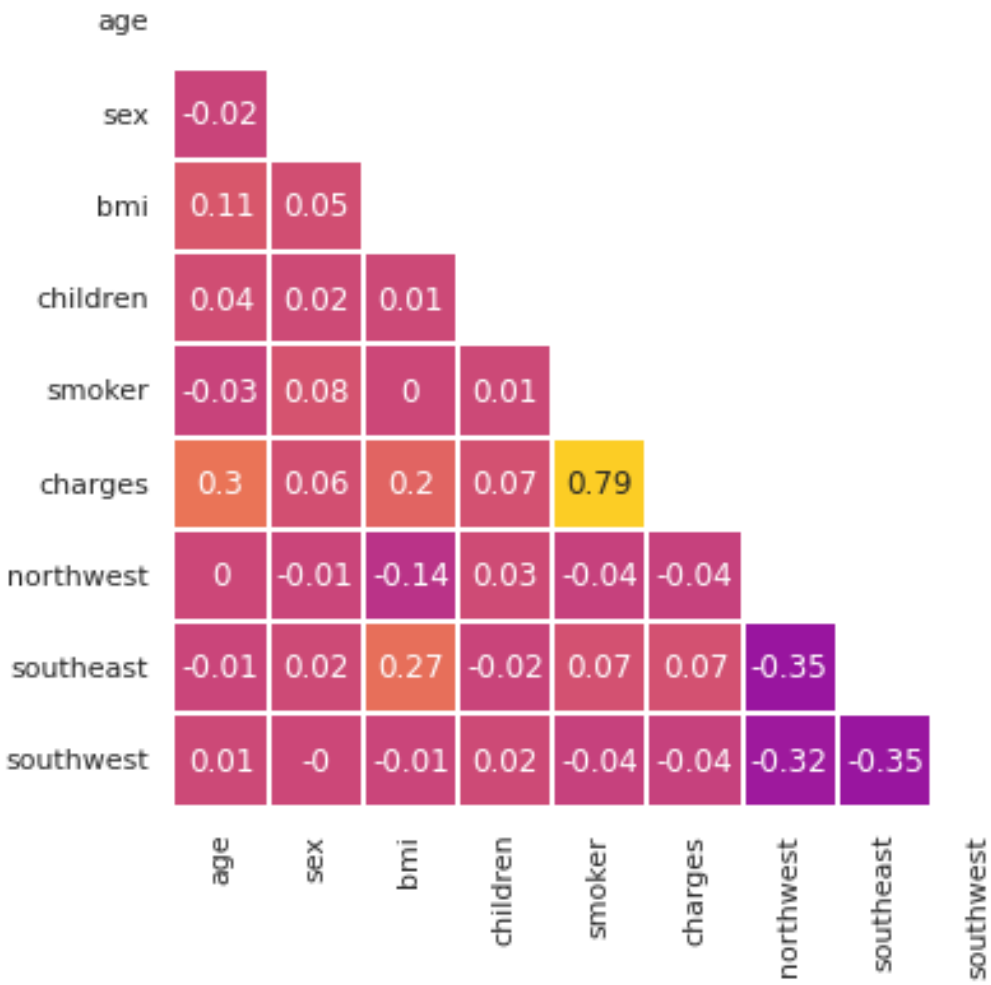

Figure 3. Pair Grid Plot

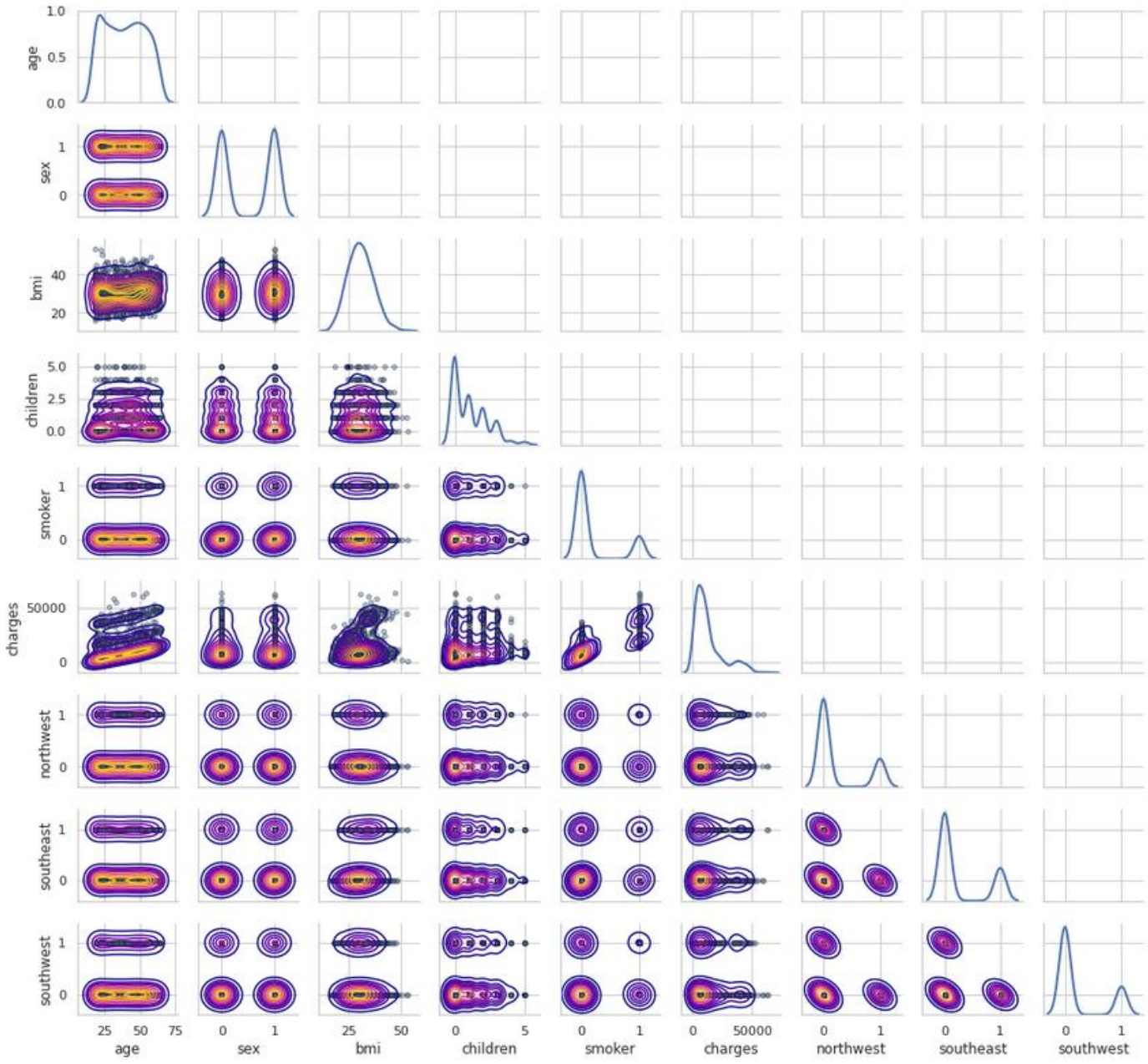

Figure 4. Digitization of attributes

3.4. Statistical significance of parameters

In order to determine the statistical significance of each feature, an Ordinary Least Squares (OLS) regression model was fitted to the data. The skewness and kurtosis have been checked and analyzed by Ch. Anwar ul Hassan et al [26]. The $p-$values of $t-$statistic were then extracted from the summary of the model to assess the significance of each feature in predicting the target variable.

Based on the threshold of $p > 0.05$ for feature significance, we dropped the 'sex' and the region-specific features ('northwest', 'southeast', 'southwest') as well. The overall model was deemed significant based on the threshold of $p > 0.05$ for F-statistic. A number of important parameters can be noticed from Table 1. Multi-co-linearity was tested using VIF and the model was tested for auto-correlation with JB-test (cf. Table 2).

Table 1. Results of the Ordinary Least Squares Model

Results: Ordinary least squares

Model:	OLS	Adj. R-squared:	0.749
Dependent Variable:	charges	AIC:	27094.2154
Date:	2023-08-29 17:11	BIC:	27140.9991
No. Observations:	1337	Log-Likelihood:	-13538.
Df Model:	8	F-statistic:	500.0
Df Residuals:	1328	Prob (F-statistic):	0.00
R-squared:	0.751	Scale:	3.6776e+07

	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
const	-11936.5575	988.2274	-12.0788	0.0000	-13875.2146	-9997.9004
age	256.7646	11.9122	21.5547	0.0000	233.3958	280.1334
sex	-129.4815	333.1952	-0.3886	0.6976	-783.1278	524.1648
bmi	339.2504	28.6113	11.8572	0.0000	283.1222	395.3786
children	474.8205	137.8969	3.4433	0.0006	204.3009	745.3401
smoker	23847.3288	413.3479	57.6931	0.0000	23036.4428	24658.2149
northwest	-349.2265	476.8238	-0.7324	0.4641	-1284.6365	586.1835
southeast	-1035.2656	478.8670	-2.1619	0.0308	-1974.6838	-95.8474
southwest	-960.0814	478.1059	-2.0081	0.0448	-1898.0066	-22.1561

Omnibus:	299.816	Durbin-Watson:	2.089
Prob(Omnibus):	0.000	Jarque-Bera (JB):	716.552
Skew:	1.211	Prob(JB):	0.000
Kurtosis:	5.646	Condition No.:	311

Table 2. Attributes versus VIF

Features	VIF
Const	35.504594
Age	1.016794
Bmi	1.106742
Children	1.004017
Smoker	1.012100
Northwest	1.517673
Southeast	1.651779
Southwest	1.529044

Table 3. Comparison of the performance of the tested Models

ML Models	MAE	RMSE	Adjusted R^2
Linear Regression	0.346	0.244	0.801
Ridge Regression	0.346	0.495	0.804
Lasso Regression	0.391	0.568	0.738
Bayesian Regression	0.346	0.495	0.801
Polynomial Regression	0.221	0.377	0.885
XGBoost	0.184	0.349	0.901
AdaBoost	0.388	0.446	0.839
Histogram-based Gradient Boosting Regression Tree	0.214	0.365	0.892
Random Forest Regressor	0.217	0.388	0.878
Decision Tree	0.253	0.509	0.790
Support Vector Regression	0.175	0.369	0.890
K-Nearest Neighbors	0.223	0.378	0.884

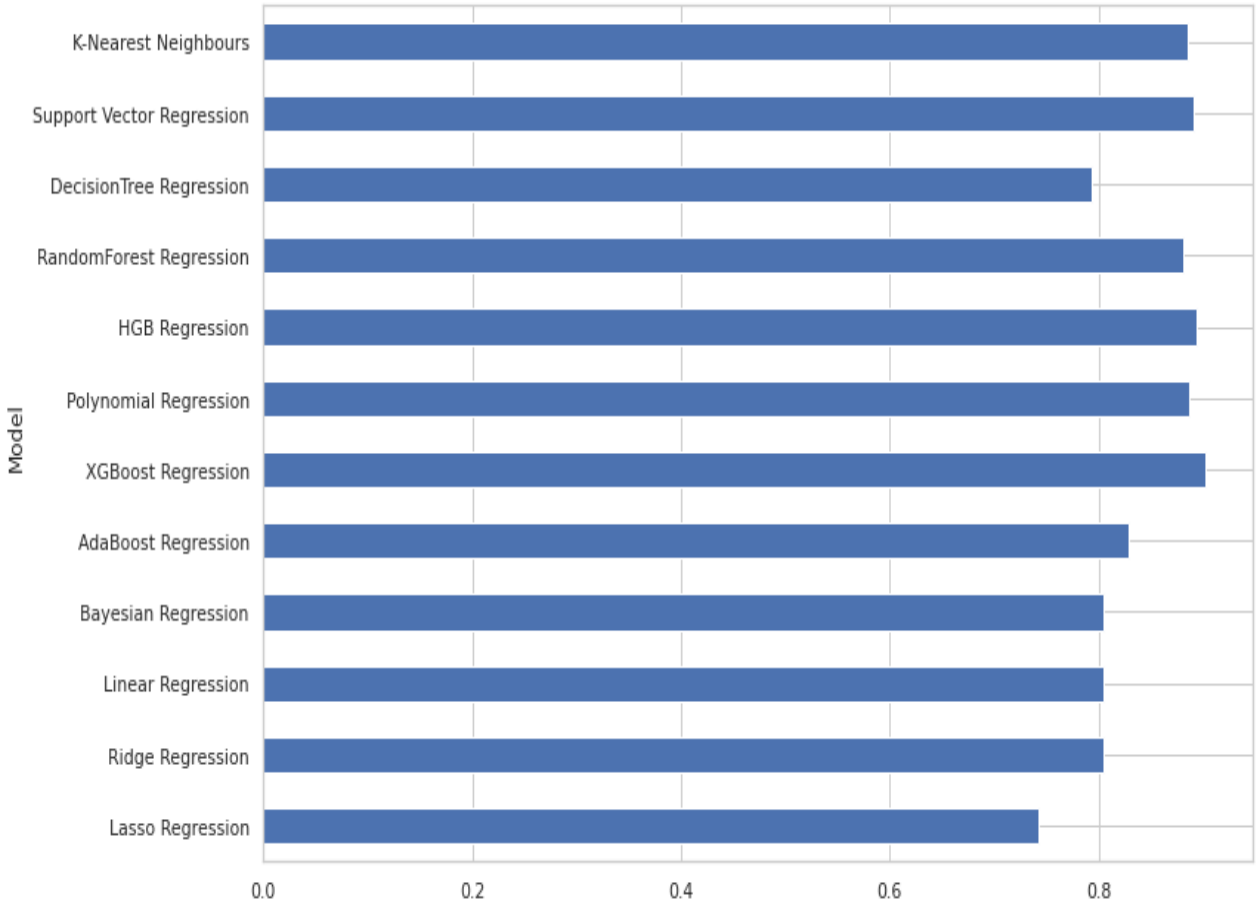

Figure 5. Graphic comparison of the Tested Models

By matching the R^2 score of these ML models in Table 3 and the bar diagram in Figure 5, it is understandable that the performance of XGBoost is the best with an adjusted R^2 of 90.

3.6. Artificial Neural Network Regression Model

The dataset was split into 20% testing and 80% training data and was further subdivided into training and validation sets. The artificial neural network model (ANN model) was constructed by using Tensor Flow's Keras API highlighting five dense layers with Exponential Linear Unit (ELU) activation functions. Dropout layers were incorporated for regularization. The model was optimized using the Adam optimizer. The model was trained for 100 epochs with a batch size of 8 and a validation split of 20%. The model accomplished a test accuracy of 93.07% and R^2 score of 92.25%.

4. Results and Analysis

The present study systematically investigates the efficacy of machine learning models in forecasting health insurance premiums by targeting enhancements in the precision of pricing for medical insurance providers. The research focused on capturing the multifaceted aspects affecting the insurance premiums such as age, bmi, smoking habits, and geographical location and put to use a diverse set of machine learning algorithms like Linear Regression, Ridge Regressor, Support Vector Regression, Random Forest and Artificial Neural Networks. A thorough study of the upshots presented in Table $1-5$ and Figure $1-5$ reveal that the results are encouraging which, in turn, demonstrates the transformative capacity of machine learning techniques in streamlining the prediction and evaluation of health insurance premiums. Peculiarly, XGBoost excelled other models with an adjusted R^2 of 90.1% and was closely trailed by the author's Artificial Neural Network model which achieved an accuracy of 93.07%. These end results affirm the central hypothesis that machine learning offers a robust and efficient methodology for accurate premium pricing and the advantage of the use of these technological tools goes directly to both the insurers and policyholders.

5. Conclusion

The research work dealt with the effectiveness of machine learning in forecasting and comparing health insurance premiums. Utilizing an open-source dataset of US medical premiums, the authors have developed a thorough machine learning model in Python programming language. Age, bmi, smoking habits, and geographical location were the main factors in the model. The results are encouraging demonstrating the transformative capacity of machine learning techniques in streamlining the prediction and evaluation of health insurance premiums. Remarkably, XGBoost outperformed other models with an adjusted R^2 of 90.1% which was closely followed by the Artificial Neural Network model of the authors which achieved an accuracy of 93.07%. Future work will explore the incorporation of nature-inspired and meta-heuristic algorithms to further refine machine learning models.

5. References

1. Your total costs for health care: Premium, deductible, and out-of-pocket costs. (n.d.). HealthCare.gov. https://www.healthcare.gov/choose-a-plan/your-total-costs/
2. Age and Sex | CMS. (n.d.). https://www.cms.gov/research-statistics-data-and-systems/statistics-trends-and-reports/nationalhealthexpenddata/age-and-gender
3. Cevolini, A.; Esposito, E. From Pool to Profile: Social Consequences of Algorithmic Prediction in Insurance. Big Data Soc. 2020, 7.
4. van den Broek-Altenburg, E.M.; Atherly, A.J. Using Social Media to Identify Consumers' Sentiments towards Attributes of Health Insurance during Enrollment Season. Appl. Sci. 2019, 9, 2035.
5. Hanafy, M.; Mahmoud, O.M.A. Predict Health Insurance Cost by Using Machine Learning and DNN Regression Models. Int. J. Innov. Technol. Explor. Eng. 2021, 10, 137–143.

6. Bhardwaj, N.; Anand, R. Health Insurance Amount Prediction. Int. J. Eng. Res. 2020, 9, 1008–1011.
7. Boodhun, N.; Jayabalan, M. Risk Prediction in Life Insurance Industry Using Supervised Learning Algorithms. Complex Intell. Syst. 2018, 4, 145–154.
8. World Health Organization: WHO. (2021, June 28). WHO issues first global report on Artificial Intelligence (AI) in health and six guiding principles for its design and use who.int.https://www.who.int/news/item/28-06-2021-who-issues-first-global-report-on-ai-in-health-and-six-guiding-principles-for-its-design-and-use
9. Goundar, S.; Prakash, S.; Sadal, P.; Bhardwaj, A. Health Insurance Claim Prediction Using Artificial Neural Networks. Int. J. Syst. Dyn. Appl. 2020, 9, 40–57.
10. Health Insurance Premium Prediction with Machine Learning. Available online: https://thecleverprogrammer.com/2021/10/26/health-insurance-premium-prediction-with-machine-learning/ (retrieved on 9 May 2022).
11. Ejiyi, C.J.; Qin, Z.; Salako, A.A.; Happy, M.N.; Nneji, G.U.; Ukwuoma, C.C.; Chikwendu, I.A.; Gen, J. Comparative Analysis of Building Insurance Prediction Using Some Machine Learning Algorithms. Int. J. Interact. Multimed. Artif. Intell. 2022,7,75-85.
12. Fauzan, M.A.; Murfi, H. The Accuracy of XGBoost for Insurance Claim Prediction. Int. J. Adv. Soft Comput. Appl. 2018, 10, 159–171 https://www.claimsjournal.com/news/national/2013/11/21/240353.htm (retrieved on 9 May 2022).
13. Rustam, Z.; Yaurita, F. Insolvency Prediction in Insurance Companies Using Support Vector Machines and Fuzzy Kernel C-Means. J. Phys. Conf. Ser. 2018, 1028, 012118.
14. Pereira, J., & Díaz, O. (2019). Using health chatbots for behavior change: A mapping study. Journal of Medical Systems, 43(5). https://doi.org/10.1007/s10916-019-1237-1
15. Kumar Sharma, D.; Sharma, A. Prediction of Health Insurance Emergency Using Multiple Linear Regression Technique. Eur. J. Mol. Clin. Med. 2020, 7, 98–105.
16. Rukhsar, L.; Bangyal, W.H.; Nisar, K.; Nisar, S. Prediction of Insurance Fraud Detection Using Machine Learning Algorithms. Mehran Univ. Res. J. Eng. Technol. 2022, 41, 33-40. https://search.informit.org/doi/epdf/10.3316/informit.2631 47785515876 (retrieved on 9 May 2022).
17. AI in the insurance sector. (n.d.). Insurance Europe. Retrieved August 15, 2023, from https://www.insuranceeurope.eu/publications/2608/artificial-intelligence-ai-in-the-insurance-sector/
18. Azzone, M.; Barucci, E.; Giuffra Moncayo, G.; Marazzina, D. A Machine Learning Model for Lapse Prediction in Life Insurance Contracts. Expert Syst. Appl. 2022, 191, 116261.
19. Sun, J.J. Identification and Prediction of Factors Impact America Health Insurance Premium. Master's Thesis, National College of Ireland, Dublin, Ireland, 2020. Available online: http://norma.ncirl.ie/4373/ (retrieved on 9 May 2022).

20. Lui, E. Employer Health Insurance Premium Prediction. Available online: http://cs229.stanford.edu/proj2012/Lui.EmployerHealthInsurancePremiumPrediction. (retrieved on 17 May 2022).
21. Prediction of Health Expense—Predict Health Expense Data. Available online: https://www.analyticsvidhya.com/blog/2021/0 5/prediction-of-health-expense/ (retrieved on 9 May 2022).
22. Yang, C.; Delcher, C.; Shenkman, E.; Ranka, S. Machine Learning Approaches for Predicting High Cost High Need Patient Expenditures in Health Care. Biomed. Eng. Online 2018, 17, 131.
23. Shyamala Devi, M.; Swathi, P.; Purushotham Reddy, M.; Deepak Varma, V.; Praveen Kumar Reddy, A.; Vivekanandan, S.; Moorthy, P. Linear and Ensembling Regression Based Health Cost Insurance Prediction Using Machine Learning. Smart Innov. Syst. Technol. 2021, 224, 495–503.
24. Transforming healthcare with AI: The impact on the workforce and organizations. (2020, March 10). McKinsey & Company. https://www.mckinsey.com/industries/healthcare/our-insights/transforming-healthcare-with-ai
25. Kaushik, K., Bhardwaj, A., Dwivedi, A. D., & Singh, R. (2022). Machine Learning-Based Regression Framework to predict health insurance premiums. International Journal of Environmental Research and Public Health, 19(13), 7898. https://doi.org/10.3390/ijerph19137898
26. Hassan, C. a. U., Iqbal, J., Hussain, S., AlSalman, H., Mosleh, M. a. A., & Ullah, S.S. (2021). A computational intelligence approach for predicting medical insurance cost. Mathematical Problems in Engineering, 2021, 1–13. https://doi.org/10.1155/2021/1162553
27. Takeshima, T., Keino, S., Aoki, Matsui, T., & Iwasaki, K. (2018). Development of medical cost prediction model based on statistical machine learning using health insurance claims data. Value in Health, 21, S97. https://doi.org/10.1016/j.jval.2018.07.738
28. Hu, S., O'Hagan, A., Sweeney, J., & Ghahramani, M. (2020). A spatial machine learning model for analyzing customer's lapse behavior in life insurance. Annals of Actuarial Science, 1–27. https://doi.org/10.1017/s1748499520000329
29. Kafuria, A. D. (2022). Predictive Model for Computing Health Insurance Premium Rates Using Machine Learning Algorithms. International Journal of Computer, 44(1), 21- 38
30. Christobel, Y. A. & Subramanian, S. (2022). An empirical study of machine learning regression models to predict health insurance cost. Webology, 19(2)