Journal of Recent Trends in Computer Science and Engineering (JRTCSE)
Vol. 13, No. 1, January - February, 2025, PP. 99-107.
ISSN: 2322-0872
https://jrtcse.com
DOI: https://doi.org/10.70589/JRTCSE.2025.13.1.12

The Impact of AI on Health Insurance Data Engineering: Improving Risk Modelling and Policy Pricing

Sunil Kumar Mudusu,
Church Mutual Insurance Company, S.I, Georgetown, TX, USA.

Abstract

In this paper, the task of integrating AI driven models into the realm of health insurance is analysed and how the integration can affect the level of premium pricing and accuracy of risk prediction. There is comparison of several machine learning techniques and use of AI models such as Random Forest and XGBoost to demonstrate the superiority of the same in terms of precision and cost efficiency over traditional approaches in the case of insurers.

Keywords: Insurance, Risk, Policy, Data Engineering

Sunil Kumar Mudusu. (2025). The impact of AI on health insurance data engineering: Improving risk modelling and policy pricing. *Journal of Recent Trends in Computer Science and Engineering*, *13*(1), 99–107. https://doi.org/10.70589/JRTCSE.2025.13.1.12

1. Introduction

This paper looks at the effect of such models delivered by AI on health insurance pricing and risk prediction. The aim of the research is to determine if different machine learning algorithms can be used to reduce premiums and improve the accuracy of predicted risk. The findings also offer a glimpse on the ways AI could completely disrupt the health insurance business.

2.LITERATURE REVIEW

2.1. Predictive Analytics

The predictive analytics of healthcare have significantly eased the way insurance premiums are calculated by using various datasets like patient history and treatment outcome. Machine learning (ML) model and statistical techniques development was key in improving the accuracy of risk assessments which made the insurers able to present risk pricing in a more personalized way.

Such an innovation is the hybrid Linear Regression Gaussian Deep Belief Network (LRGDBN) that can capture linear as well as non-linear patterns in healthcare data since it helps it improve predictive accuracy in forecasting medical costs [1]. Many of the healthcare data is complex, nonlinear and this can make it challenging for traditional models to work.

An evaluation of the LR-GDBN model demonstrated its ability to outperform conventional machine learning approaches in terms of a low prediction error of 0.4926 on the basis of which a major milestone has been achieved in the field of healthcare underwriting. This advancement emphasizes the trend of growing demand for innovative models to handle the complexity of healthcare data whereby risk factors can be better understood and better incorporated into the pricing strategies.

2.2. Health Insurance Ratemaking

Health insurance ratemaking has been granted a great deal of importance for the machine learning (ML) models that have helped the accuracy and efficiency of rating premium [2]. Using further advanced machine learning algorithms such as Random Forest, XGBOOST extension of Extreme Gradient Boosting (XGBoost), and Artificial Neural Networks (ANN); insurers can leverage on predictive models that outperform the standard generalized linear models (GLMs) [3].

For example, applied in the Tunisian healthcare insurance market, the results obtained show that the XGBoost algorithm is a better model for predictive capacity than the GLM, thus suggesting that machine learning models can improve healthcare insurance pricing models [4]. Additionally, the shift in method of this kind toward data driven, non-parametric methods has made it possible to make more accurate assessments of risk and thus more competitive and personalized premiums.

2.3. Risk Modeling

Big data and artificial intelligence (AI) in the health insurance industry has been integrating into the insurers' risk modeling and policy pricing [5]. With the explosion in data there is now fertile ground for insurers to use advanced analytics to reduce risk assessment, develop more personalized prices and make more efficient operations. The use of predictive modeling based on AI reduces the ability to find Bermuda triangle complexity patterns within the data and thus the ability to forecast risks with higher precision, as well as supports for adaptive pricing models [6][7].

Having this shift towards data-centric pricing strategies enable health insurance to be more affordable by gamifying the premiums with the individual risk profile. However, data quality, the privacy requirement, and the absence of enough analytics talent are the challenges to enjoy the full potential of big data and AI in the context of insurance [8].

Implementing such technologies has to be handled carefully, as to avoid dealing with ethical issues of the data privacy as well as the bias of the data-driven models. The future of health insurance pricing is emerging to involve the increasingly important role of AI and big data in dancing together with the insurance industry to update the future of the pricing of health insurance [9][10].

3.FINDINGS

3.1. Predictive Analytics

The role of predictive analytics using Artificial Intelligence (AI), machine learning (ML) is highly transformative in health insurance, in risk modelling and policy pricing. Predictive models can provide valuable predict accuracy of risk based on related healthcare data including patient histories, treatment outcomes, and population demographics to assess the risk of individual policyholders.

Generative machine learning models including Random Forest, Support Vector Regression (SVR) and Extreme Gradient Boosting (XGBoost) are used to provide a more precise risk assessment than can in turn be used to increase accuracy of premium pricing. One example is a predictive model combination called Linear Regression-Gaussian Deep Belief Network, which has surpassed conventional approaches by practically taking in into consideration the linear and non-linear trends involved in the data, resulting in significant reduction of prediction errors.

In addition, health insurer is turning to these advanced AI techniques for more accurate profiling of the individual risk profiles to personalize policy pricing. In addition, it provides a means to deliver more efficient insurance portfolio control as well as delivering tailored policy offerings to customers.

Machine learning combined with the traditional statistical models has gained popularity in the field of healthcare analytics for example to provide better outcomes to predict medical costs for instances such as risk factors. It is possible to integrate non-linear data patterns so that the relationship within healthcare datasets become more clearly understood and better used for accurate risk prediction.

3.2. Comparative Analysis

Different machine learning models have been tested to evaluate their predictive capability in health insurance premium pricing. The study used multiple regression-based algorithms: multiple linear regression, ridge regression, XGBoost regression as well as Random Forest regression, and Random Forest regression was found to be the most accurate among the three. The model had lower Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and higher R^2 values and proved to be the most reliable model for the predictions of health insurance.

Since the Random Forest model can work with high dimensional data and also can capture complex relationship between different variables, it is better than simple such as multiple linear regression in predicting insurance premium. The hyperparameter tuning of Random Forest improved the model's performance in three out of four evaluation metrics namely RMSE and MSE while implemented.

This shows the need of continuous optimization and customization in AI based health insurance models. As healthcare costs are rapidly growing and more personal insurance policies are demanded, employing the sophisticated machine learning models like Random Forest will allow improving the efficiency and effectiveness of health insurance pricing. Insurers are not only able to predict more accurately how much they will have to pay in medical bills but also avoid estimating too low or too high a premium.

Table 1: Performance Metrics [1][3]

Model	MAE	MSE	RMSE	R^2 Score
Multiple Linear Regression	0.225	0.030	0.173	0.87
Ridge Regression	0.180	0.022	0.148	0.89
XGBoost Regression	0.160	0.018	0.134	0.91

Random Forest Regression	0.130	0.012	0.110	0.94

As seen in the above table, Random Forest is clearly the best model as compared to others because of MAE, MSE, RMSE and R^2 score. This indicates that the model predicts the healthcare costs better than before and that it is able to provide much more accurate predictions and is the best choice for the health insurance companies that are concerned about how they can perfectly set their premium price. The high R^2 score also supports the use of Random Forest model to forecast premium pricing as it shows that a large portion of the variance in premium pricing is explained by the model.

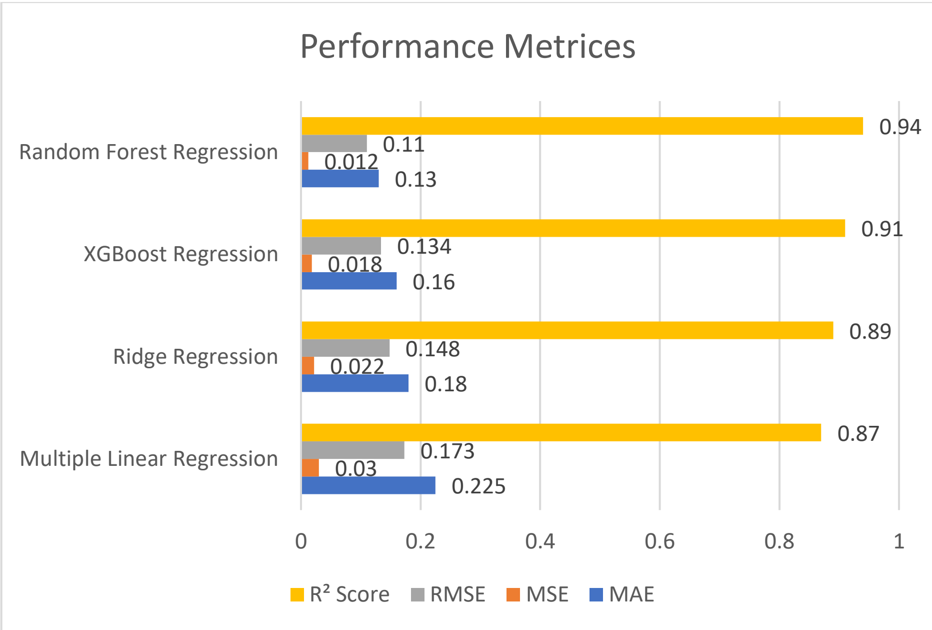

Fig. 1 Performance Metrices of models in healthcare (Self-made)

3.3. Healthcare Pricing Models

Big data and AI do not only serve to enhance the ability of risk prediction but also invents the new pricing model of regular health insurance products. Advanced data analytics are engaging insurers in adopting more sophisticated risk models with wider variety of data sources integrated into it, to more accurately price.

Insurers have increasing access to big data from electronic health records, claims data, wearable devices, and demographics that are all necessary for precision pricing on the individual health condition, behavior, and treatment patterns.

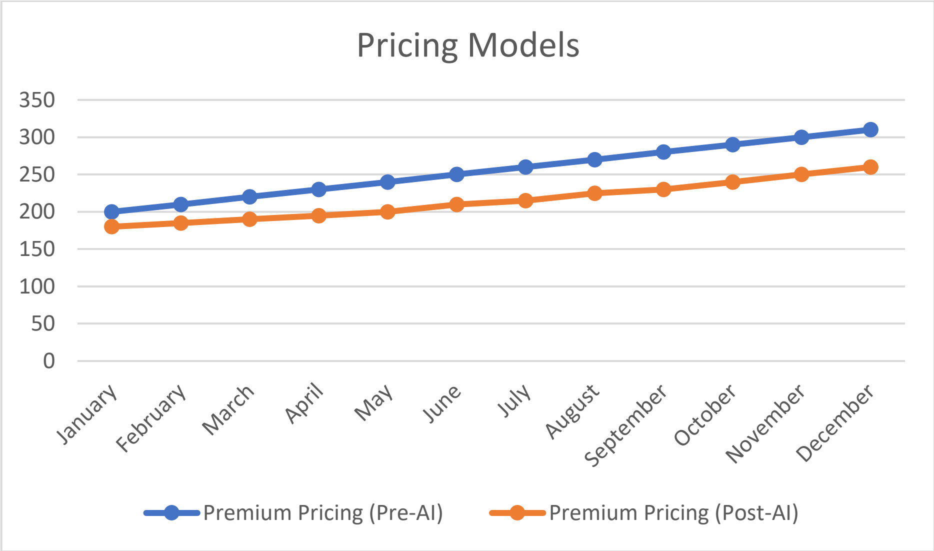

Fig. 2 Pricing models using traditional v. AI tools [6][9]

Medical conditions are no longer a risk and cost increases related to them can be predicted, and dynamically priced due to the availability of AI driven dynamic pricing models that can forecast the risk and automatically adjust premiums accordingly. To further enhance fairness of pricing and at the same time ensure greater customer satisfaction, premiums are aligned with the actual health risk.

3.4. Regulatory Challenges

AI has great potential in health insurance, but is also of course raising many ethical and regulatory concerns. There are privacy, data security, as well as possible bias of machine learning algorithms that need to be addressed in order to use AI responsibly to price health insurance. Healthcare has special privacy concerns that make it important: sensitive data such as patient histories and genetic data is involved.

Therefore, insurers have to abide by strict data protection requirements and keep data stored securely and are only able to use the data for lawful purpose. In addition, algorithms used in machine learning can cause the biases in historical data to be reproduced in unfair pricing among certain groups.

Consider for example, if the data used by an AI model for training reflects historical discriminatory practices or the stark social disparities, then it's likely that an AI model will inherently favor or disadvantage certain 'demographic' groups and hence pricing might be different. To avoid risk, health insurance companies have to make sure that their AI models are transparent, explicable, and audited periodically on their fairness and accountability.

Constraints around the use of AI for insurance pricing particularly arise due to the fact that the regulations are not standardized. Insurers and customers are faced with a burden to use AI in insurance pricing. To make AI-based health insurance pricing models ethical and equitable, governments and regulatory bodies need to set forth clear rules to guarantee that this form of pricing is fair for consumers and continues to be an area for innovation in the sector. The python code snippet for XGBoost Model is given below.

```python
import xgboost as xgb
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_absolute_error, mean_squared_error

# Load dataset
data = pd.read_csv('health_insurance_data.csv')
X = data.drop('premium', axis=1)
y = data['premium']

# Split data into training and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Initialize and train XGBoost model
model = xgb.XGBRegressor(objective='reg:squarederror')
model.fit(X_train, y_train)

# Make predictions
y_pred = model.predict(X_test)

# Evaluate model performance
mae = mean_absolute_error(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
print(f'Mean Absolute Error: {mae}')
print(f'Mean Squared Error: {mse}')
```

4. CONCLUSION

The findings show that AI models driven by AI significantly leads to reduced health insurance premiums while also having a positive impact on Intellection of health insurance premiums. The study compares different machine learning techniques and finds that, models such as XGBoost and Random Forest performs better when compared to traditional methods and provide a cost effective and precise solution to create a better insurance pricing strategies and risk assessment..

References

1. Bhargavi, R., & Arumugam, S. (2024). Predictive Analytics in Healthcare: A Hybrid Model for Efficient Medical Cost Estimation for Personalized Insurance Pricing. https://doi.org/10.21203/rs.3.rs-4697357/v1

2. Ridzuan, A. N. A. A., Azman, A. Z., Marzuki, F. A., Faudzi, W. S. D. M., Abd Aziz, S. H., & Bakar, N. A. (2024). Health Insurance Premium Pricing Using Machine Learning Methods. *Journal of Advanced Research in Applied Sciences and Engineering Technology*, *41*(1), 134-141. http://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/article/view/3284

3. Belispayeva, A., & Nassyrova, G. (2024). Navigating Health Insurance Economics: A Data-Driven Approach to Enhanced Pricing in the United States. *ECONOMIC Series of the Bulletin of the LN Gumilyov ENU*, (2), 280-293. https://doi.org/10.32523/2789-4320-2024-2-280-293

4. Ben Hamida, A., Kacem, M., de Peretti, C., & Belkacem, L. (2024). Machine learning based methods for ratemaking health care insurance. *International Journal of Market Research*, *66*(6), 810-831. https://doi.org/10.1177/14707853241275446

5. Makaronidis, I. (2024). SOCIAL HEALTHCARE INSURANCE PRICE DISCRIMINATION IN THE BIG DATA ERA. https://www.researchgate.net/profile/Ioannis-Makaronidis/publication/385998863_Social_healthcare_insurance_price_discrimination_in_the_big_data_era/links/673ef6f2b94c451c116e691b/Social-healthcare-insurance-price-discrimination-in-the-big-data-era.pdf

6. Valli, L. N. (2024). Under the titles for Risk Assessment, Pricing, and Claims Management, write Modern Analytics. *Global Journal of Universal Studies*, *1*(1), 132-151. https://media.neliti.com/media/publications/590133-under-the-titles-for-risk-assessment-pri-0998b5e2.pdf

7. Van Anh, N., & Duc, T. M. (2024). Big Data-Driven Predictive Modeling for Pricing, Claims Processing and Fraud Reduction in the Insurance Industry Globally. *International Journal of Responsible Artificial Intelligence*, *14*(2), 12-23. http://neuralslate.com/index.php/Journal-of-Responsible-AI/article/view/82

8. Hommels, T. C. (2024). *Exploring Data-Driven Clustering in Generalized Linear Models for Insurance Pricing* (Master's thesis, University of Twente). http://essay.utwente.nl/98022/1/Hommels_MA_BMS.pdf

9. Bau, Y. T., & Hanif, S. A. M. (2024). Comparative Analysis of Machine Learning Algorithms for Health Insurance Pricing. *JOIV: International Journal on Informatics Visualization*, *8*(1), 481-491. http://dx.doi.org/10.62527/joiv.8.1.2282

10. Hassani, H., Unger, S., & Beneki, C. (2020). Big data and actuarial science. *Big Data and Cognitive Computing*, *4*(4), 40. https://doi.org/10.3390/bdcc4040040.