Predictive Analytics for Health Insurance Pricing Using ML and Explainable AI

ABSTRACT

Accurate prediction of health insurance costs is essential for reducing financial risk and creating fair pricing strategies for insurers and customers. Traditional actuarial methods rely on statistical formulas and manual analysis, which often fail to capture complex interactions among health factors, lifestyle attributes, and demographics. This research presents an AI-driven prediction system that uses machine learning algorithms to estimate health insurance charges with high accuracy. Models such as Linear Regression, Random Forest, XGBoost, and Artificial Neural Networks were trained and evaluated on publicly available medical cost datasets. Performance metrics including MAE, RMSE, and R^2 score were used to assess model accuracy. Experimental results demonstrate that tree-based ensemble models outperform classical approaches by capturing nonlinear relationships effectively. The proposed system also incorporates explainable AI (SHAP) to interpret feature importance and enhance transparency in decision-making. The study highlights the potential of AI to support insurance companies, healthcare providers, and customers by providing reliable cost predictions and improving pricing efficiency.

1. INTRODUCTION

Health insurance costs vary widely depending on demographic, lifestyle, and medical factors. Predicting these costs accurately is crucial for insurers to maintain profitability while offering fair premiums. Traditional risk assessment models often ignore complex nonlinear interactions and rely heavily on manual assumptions. Artificial Intelligence (AI) and Machine Learning (ML) provide powerful tools for modeling these interactions and improving prediction accuracy. This paper investigates how ML algorithms can predict insurance costs more effectively and how explainable AI can improve transparency and trust.

2. PROBLEM STATEMENT

Current insurance pricing models:

- Depend on predefined statistical formulas
- Struggle to capture nonlinear relationships
- Lack transparency in decision-making
- Result in inconsistent premium calculations

There is a need for an AI-based system that provides accurate, fair, and interpretable cost predictions.

3. LITERATURE REVIEW

Previous studies have used:

- Linear Regression for baseline predictions
- Decision Trees for non-linear patterns
- Random Forest and Gradient Boosting for higher accuracy
- Neural networks for complex feature interactions
- SHAP/LIME for explainability

However, many studies lack full evaluation across multiple modern ML models and do not include explainability for real-world adoption.

4. Dataset Description

Example dataset:

- Medical Cost Personal Dataset
  Features include:
- Age
- BMI
- Smoking status
- Number of children
- Region
- Gender
- Charges (target variable)

You can also use enhanced datasets if available.

5. Methodology

Step 1: Data Preprocessing

- Handling missing values
- One-hot encoding for categorical fields
- Normalization/standardization
- Outlier detection

Step 2: Algorithms Used

- Linear Regression
- Random Forest
- XGBoost
- Support Vector Regression
- Neural Network (optional)

Step 3: Model Evaluation

Metrics used:

- MAE (Mean Absolute Error)
- MSE (Mean Squared Error)
- RMSE (Root Mean Square Error)
- R^2 Score

Step 4: Explainability

Use SHAP to identify:

- Which features impact cost most
- How smoking, age, and BMI affect charges

6. Experimental Results

Sample (You can fill with your actual numbers):

- Linear Regression: $R^2 = 0.75$
- Random Forest: $R^2 = 0.88$
- XGBoost: $R^2 = 0.91$
- Neural Network: $R^2 = 0.89$

XGBoost performed best, achieving the highest accuracy and lowest error.

Top predictive factors (from SHAP):

- Smoking status
- BMI
- Age
- Number of children

7. Discussion

The results show that AI models, especially ensemble learning algorithms, significantly improve prediction accuracy compared to traditional methods. SHAP visualization provides transparency, allowing insurers to understand why certain customers receive higher or lower charges. This increases fairness and trust in pricing.

8. Conclusion

This study demonstrates that AI-driven machine learning models can predict health insurance costs with high accuracy. XGBoost and Random Forest outperform conventional methods by effectively modeling nonlinear interactions. The inclusion of explainability techniques strengthens the credibility of predictions and supports ethical insurance pricing.