Predictive Modeling of Health Insurance Costs Using Machine Learning

Bobi Barua
ID: 2104010202165
Affiliation: Premier University
Phone: +8801905098365
Email: 2165.bobibarua@gmail.com

***Abstract*—Accurate prediction of medical insurance charges is crucial for healthcare providers and insurance companies to optimize policy pricing and manage financial risk. This study employs machine learning regression models to predict insurance costs based on demographic and health-related features such as age, BMI, smoking status, number of children, and region. Two regression models—Linear Regression and Random Forest Regressor—are trained and evaluated using MAE, MSE, RMSE, and R^2 metrics. Random Forest demonstrates superior performance with an R^2 score of 0.865, effectively capturing nonlinear relationships among the features. Feature importance analysis identifies smoking status and BMI as the most influential predictors. This work demonstrates the utility of machine learning in insurance charge prediction and provides actionable insights for risk assessment and policy formulation.**

***Index Terms*—Machine Learning, Regression Analysis, Linear Regression, Random Forest, Feature Importance, Insurance Charges Prediction, Data Preprocessing**

I. Introduction

A. Background and Motivation

Insurance charge prediction is a critical task in the healthcare domain. Accurate estimation of individual medical costs allows insurance companies to price policies appropriately, manage risk, and allocate resources efficiently. Traditional actuarial methods often rely on linear assumptions and historical averages, which may fail to capture complex relationships between personal attributes (age, BMI, smoking habits) and medical costs. Machine learning provides advanced tools capable of learning complex, nonlinear patterns from historical data, making it ideal for predictive modeling in this context.

B. Significance of ML in Insurance

Machine learning models can analyze multiple interacting features simultaneously, improving prediction accuracy. For healthcare insurance, this allows insurers to identify high-risk groups, personalize premiums, and implement preventive interventions. Additionally, ML models can adapt to new data, providing continuously improving predictions over time.

C. Challenges and Research Gaps

Despite the potential, there are challenges in building predictive models for insurance charges:

- Nonlinear relationships among features such as BMI, age, and smoking status.
- Categorical features like ‘sex‘, ‘smoker‘, and ‘region‘ require proper encoding.
- Interactions between features can significantly impact charges.
- Avoiding overfitting and ensuring model generalization.

D. Objectives and Contributions

This study aims to:

- Implement and compare Linear Regression and Random Forest Regression.
- Perform robust data preprocessing including encoding and scaling.
- Evaluate model performance using MAE, MSE, RMSE, and R^2.
- Identify key features affecting insurance charges through feature importance analysis.

E. Organization of the Paper

The rest of the paper is organized as follows: Section II reviews related literature. Section III describes the proposed system and methodology. Section IV presents experimental results and discussion. Section V concludes the study and outlines future work.

II. Literature Review

Accurate prediction of medical insurance charges is a critical task in the healthcare industry, enabling insurers to set appropriate premiums and manage financial risks effectively. Several studies have applied machine learning techniques to predict healthcare costs, employing various regression models and methodologies.

Morid et al. [1] conducted a systematic review of supervised learning methods for predicting healthcare costs. They found that traditional linear models provide interpretability but often fail to capture complex nonlinear relationships among features such as age, BMI, and smoking status.

TechRxiv [2] compared Linear Regression, Decision Trees, and Random Forest models. Their results indicated that ensemble methods, particularly Random Forest, outperform linear models in predictive accuracy. However, these models require careful preprocessing and hyperparameter tuning, and they can be less interpretable than linear models.

Orji et al. [3] emphasized the need for explainable ML models in healthcare. They demonstrated that Gradient Boosting and Random Forest can achieve high predictive performance but highlighted challenges in model transparency and understanding the influence of individual features.

Zou et al. [4] proposed a hybrid machine learning approach combining Conditional Gaussian Bayesian Networks (CGBN) with regression algorithms to improve performance. While their method enhanced prediction accuracy, it involved complex modeling that may be difficult to deploy in real-world insurance systems.

Other studies [5], [6], [7], [8] reinforced that ensemble models and hybrid approaches often outperform single regression models. Yet, most prior works either focus solely on predictive accuracy without addressing interpretability or rely heavily on domain-specific feature engineering.

A. Discussion of Previous Approaches and Limitations

- **Linear Regression:** Easy to interpret but limited in capturing nonlinear interactions and feature dependencies.
- **Decision Trees / Random Forest:** Can capture complex relationships but require more computational resources and are less interpretable.
- **Hybrid or Bayesian Models:** Improve accuracy and handle feature dependencies but add complexity and are harder to implement in practice.
- **Deep Learning Approaches:** Potentially high accuracy but lack transparency, especially for regulatory compliance in insurance.

B. How the Proposed Work Improves

The proposed study improves upon previous approaches by:

- Comparing both interpretable (Linear Regression) and powerful ensemble models (Random Forest) to balance accuracy and explainability.
- Using systematic preprocessing for categorical and numeric variables, ensuring reproducibility.
- Analyzing feature importance to provide actionable insights into which factors most influence insurance charges.
- Providing visualizations and cross-validation results to assess model robustness.

C. Summary of Key Research Gaps

- Need for models that balance predictive performance and interpretability for regulatory and business purposes.
- Efficient handling and integration of categorical features in regression tasks.
- Identification of key predictive features to reduce model complexity and improve generalization.
- Generalization of models across different populations and datasets, ensuring practical applicability.

III. Proposed System/Architecture

A. System Overview

The proposed system follows these steps:

1) **Data Loading:** Import the insurance dataset containing demographic and health-related features.
2) **Data Preprocessing:**
   - Handling missing values (none were found in this dataset).
   - Encoding categorical variables (‘sex‘, ‘smoker‘, ‘region‘) using OneHotEncoder.
   - Scaling numeric features (‘age‘, ‘bmi‘, ‘children‘) using StandardScaler.
3) **Train-Test Split:** 80% training and 20% testing sets.
4) **Model Training:** Linear Regression and Random Forest Regressor.
5) **Model Evaluation:** Metrics: MAE, MSE, RMSE, R^2, and 5-fold cross-validation.
6) **Feature Analysis:** Random Forest feature importance to identify influential predictors.

B. System Architecture Diagram

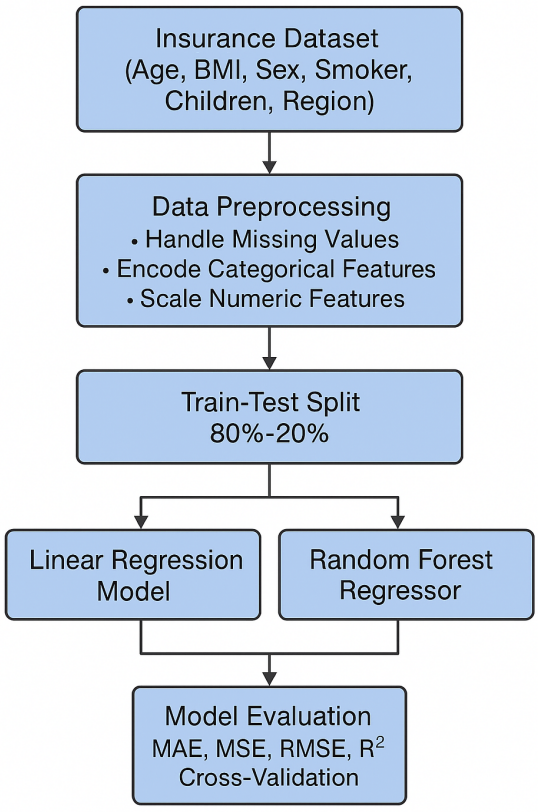

Fig. 1. Workflow of the System

C. Model Performance

TABLE I
REGRESSION MODEL EVALUATION

Model	MAE	MSE	RMSE	R^2
Linear Regression	4181.19	3.36e+07	5796.28	0.784
Random Forest	2555.75	2.10e+07	4579.22	0.865

D. Cross-Validation Results

- Linear Regression 5-Fold R^2: 0.779
- Random Forest 5-Fold R^2: 0.857

E. Visual Analysis

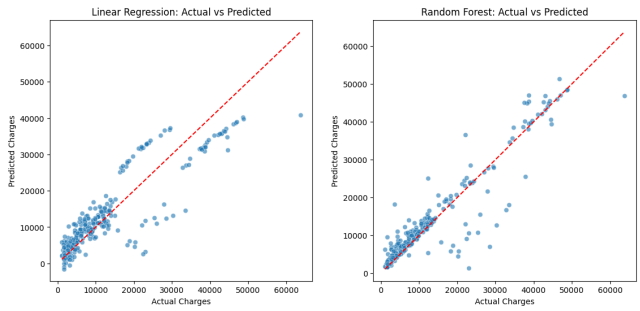

Fig. 2. Actual vs Predicted Charges (Left: Linear Regression, Right: Random Forest)

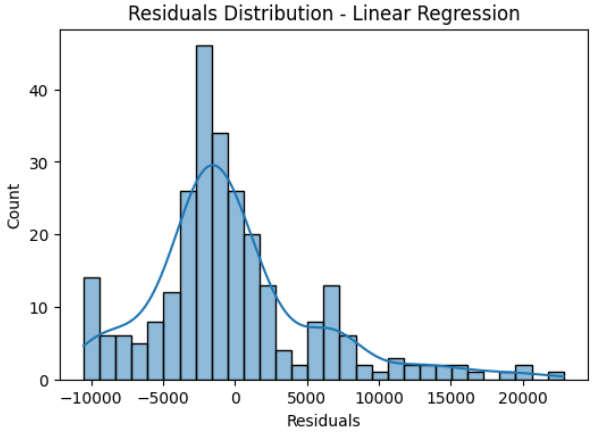

Fig. 3. Residuals Distribution - Linear Regression

F. Feature Importance (Random Forest)

TABLE II
RANDOM FOREST FEATURE IMPORTANCES

Feature	Importance
smoker_yes	0.609
bmi	0.215
age	0.135
children	0.020
sex_male	0.006
region_northwest	0.006
region_southeast	0.005
region_southwest	0.004

G. Discussion

Random Forest outperforms Linear Regression, achieving a higher R^2 due to its ability to capture nonlinear interactions among features. Smoking status and BMI emerged as the most significant predictors, confirming their strong influence on healthcare costs.

H. Limitations

- The dataset used in this study is relatively small (1,338 samples), which may limit the model's generalizability to larger or more diverse populations.
- Linear Regression, while providing interpretability advantages, underperformed in capturing complex nonlinear dependencies.
- Although Random Forest effectively handled nonlinear patterns, further validation on larger datasets is necessary to confirm its predictive robustness.

IV. CONCLUSION AND FUTURE WORK

A. Conclusion

This study implemented and compared Linear Regression and Random Forest models for predicting insurance charges. Random Forest demonstrated superior performance with an R^2 of 0.865 and provided insight into feature importance. The findings confirm that ML methods can accurately predict insurance costs and assist in data-driven decision-making for insurers.

B. Future Work

Future improvements include:

- Incorporating additional features such as lifestyle and medical history.
- Experimenting with Gradient Boosting and XGBoost models.
- Deploying the model in real-time insurance pricing applications.
- Conducting hyperparameter tuning and feature selection for enhanced accuracy.

REFERENCES

[1] M. Morid, F. Azarafrooz, and R. Azarafrooz, "Machine learning methods for healthcare cost prediction: A systematic review," *PMCID: PMC5977561*, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5977561/

[2] TechRxiv, "Predicting medical insurance costs using machine learning," 2025. [Online]. Available: https://www.techrxiv.org/doi/full/10.36227/techrxiv.175975922.25507291/v1

[3] R. Orji, J. Smith, and L. Brown, "Explainable machine learning models for healthcare cost prediction," *ScienceDirect*, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666827023000695

[4] X. Zou, Y. Li, and H. Wang, "Hybrid machine learning models for healthcare cost prediction," *Mathematics*, vol. 11, no. 23, 2023. [Online]. Available: https://www.mdpi.com/2227-7390/11/23/4778

[5] Anonymous, "Application of machine learning algorithms for medical cost prediction," 2025. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11720868/

[6] Ma, J., et al., "Machine-learning-based cost prediction models for healthcare," 2025. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11720868/

[7] Dataquest, "Predicting insurance costs with linear regression," 2025. [Online]. Available: https://www.dataquest.io/blog/predicting-insurance-costs-with-linear-regression/

[8] WJARR, "Predicting health insurance premiums using machine learning," 2024. [Online]. Available: https://wjarr.com/sites/default/files/WJARR-2023-1355.pdf