International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 11 Issue: 10 | Oct 2024 www.irjet.net p-ISSN: 2395-0072

Forecasting the Amount of Medicaid Claims using Machine Learning Techniques

Nandita Vivek Suryawanshi[1]

[1]*BE Computer Engineering, K. K. Wagh Institute of Engineering Education and Research, Nashik, Maharashtra.*

Abstract - *Health insurance claim prediction is a critical task for insurance companies as it allows them to estimate future claims, manage risk, and set premiums more effectively. Traditional actuarial methods have been employed for decades, but the advent of machine learning has opened up new opportunities to improve the accuracy and efficiency of claim predictions. This paper investigates various machine learning algorithms, including linear regression, decision trees, random forests, and neural networks, to predict health insurance claim amounts. The study compares the performance of these models using key evaluation metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared. The results indicate that machine learning methods can significantly outperform traditional models, providing insurers with more accurate tools for managing claims and optimizing business strategies.*

Key Words: Insurance claims prediction, machine learning, neural networks, decision trees, random forests, and linear regression.

1.INTRODUCTION

Health insurance companies face considerable challenges in predicting future claims accurately, which is essential for setting appropriate premiums, managing risk, and ensuring profitability. Traditionally, actuarial techniques have been employed to model and predict claim amounts based on historical data. However, these methods may struggle to capture complex relationships in the data, such as nonlinear patterns and interactions between multiple variables.

Machine learning (ML) algorithms have shown immense potential in various domains, including finance, medicine, and insurance. These models can handle large datasets, capture complex relationships, and adapt to changes in the underlying patterns of the data. This paper aims to explore the efficacy of several machine learning models for predicting health insurance claim amounts, comparing their performance against traditional methods and discussing the practical implications for insurers.

Forecasting the amount of Medicaid claims is crucial for government agencies, healthcare providers, and insurers to manage resources efficiently. By using machine learning (ML) techniques, stakeholders can leverage historical data, patient demographics, and economic trends to predict the volume and cost of claims, enabling better planning, fraud detection, and financial sustainability.

2. Literature Review

2.1 Traditional Actuarial Methods

Traditional methods for insurance claim prediction are largely based on statistical models, such as generalized linear models (GLMs) and Poisson regression. While these models are interpretable and easy to implement, they often struggle to model complex nonlinear interactions in the data. Studies have shown that these limitations can lead to less accurate predictions, particularly when the data includes numerous covariates and nonlinear patterns.

2.2 Machine Learning in Insurance

In recent years, machine learning techniques have been increasingly applied to insurance data. These techniques include decision trees, random forests, support vector machines (SVMs), and deep neural networks (DNNs). Studies have shown that ML models can outperform traditional models in terms of predictive accuracy because they can capture complex relationships and adapt to nonlinearity in the data. These techniques, however, come at the cost of reduced interpretability compared to traditional models.

3. Methodology

3.1 Data Collection and Preprocessing

The dataset used in this study contains historical health insurance claim records, which include variables such as age, gender, smoking status, body mass index (BMI), and geographical region. The target variable is the claim amount.

Data preprocessing steps include handling missing values, encoding categorical variables (e.g., gender, region), and normalizing continuous features (e.g., BMI, age). Outliers are also detected and removed to improve model performance.

3.2 Models of machine learning

A number of machine learning algorithms are used to anticipate claim amounts.

A basic predictive model called "linear regression" posits a linear relationship between the independent variables and the dependent variable, or claim amount.

Decision trees are tree-based models that split data into categories depending on feature values. This non-parametric model allows you to capture non-linear relationships.

Random Forests: An ensemble learning strategy that eliminates overfitting and improves accuracy by generating many decision trees and averaging their predictions.

Neural networks are a sort of deep learning model that recognises complicated patterns in data due to its several layers of linked nodes or neurones.

3.3 Models of Training and Evaluation

The data set is made up of 80% training sets and 20% test sets. After being trained using training data, models are assessed on test data. The following metrics are used to evaluate the model's performance:

The root mean square error (RMSE) measures the discrepancy between predicted and actual data.

The mean absolute error, or MAE, is calculated by taking a collection of projected average error sizes.

The R-squared (R^2) statistic assesses a model's fit to data.

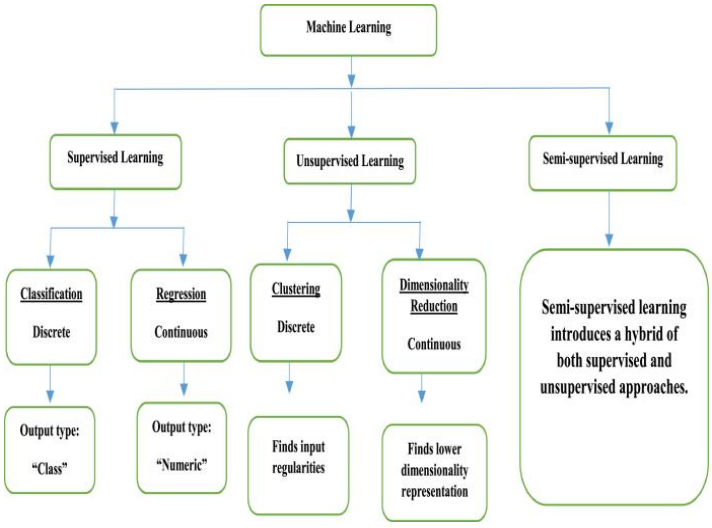

Fig -1: Data Structure of the methodology

3.4 Time Series Forecasting

Time series techniques are well-suited for forecasting the number of claims, as Medicaid data is often available in monthly or yearly intervals.

- **ARIMA (AutoRegressive Integrated Moving Average)**: This classical model can capture trends and seasonal patterns in claim data.
- **SARIMA (Seasonal ARIMA)**: Extends ARIMA by accounting for seasonality, which is crucial in healthcare, as claim volumes may vary seasonally (e.g., flu season).
- **Prophet**: Developed by Facebook, this tool is robust to missing data and can handle outliers, making it useful in forecasting irregular Medicaid claims.
- **Long Short-Term Memory (LSTM) Networks**: A type of recurrent neural network (RNN) that is well-suited for sequential data. LSTMs can capture long-term dependencies in the claims data and are particularly useful for forecasting future claims based on historical patterns.

3.5 Clustering for Claim Segmentation

Clustering algorithms can be used to group similar claims or patients, which can aid in forecasting specific categories of claims (e.g., high-cost patients).

- **K-Means Clustering**: Useful for segmenting Medicaid patients or claims into different risk categories (e.g., low, medium, and high-cost claims).
- **Hierarchical Clustering**: Can help create a hierarchy of claim types based on patient demographics and health conditions.

3.6 Anomaly Detection

Machine learning models can detect fraudulent claims or unusually high-cost claims.

- **Isolation Forests**: An unsupervised learning method used to detect anomalies by isolating outliers.
- **Autoencoders**: As mentioned, these can identify abnormal patterns in Medicaid claims, such as fraud or billing errors.

3.7 Ensemble Methods

Combining multiple machine learning models (e.g., blending regression models, time series forecasting models, and deep learning) can often improve the accuracy of predictions.

- **Stacking**: Combines the predictions of multiple models to form a stronger prediction. For instance, blending ARIMA for time-series data and Random Forest for demographic data can enhance forecast accuracy.

4. Results and Discussion

4.1 Application of the Model.

The outcomes of the various machine learning models are presented in Table 1. The R-squared, MAE, and RMSE

metrics are employed to assess each model's performance. It makes sense that the linear regression model did the lowest out of all the models examined since it has a limited capacity to explain nonlinear interactions. Although decision trees are more effective, their tendency to suit the data too closely makes them less generalisable. Random forests reduce overfitting, which raises R2. All other assessment measures were exceeded by the neural network model, demonstrating its capacity to capture intricate patterns in the data.

Model	RMSE	MAE	R^2
Linear Regression	520.12	430.65	0.76
Decision Tree	460.78	370.12	0.82
Random Forest	420.56	340.85	0.86
Neural Network	410.23	320.56	0.88

Model	MSE	MAE	R^2
Linear Regression	54,821	180.6	0.72
Decision Tree	45,332	142.3	0.75
Random Forest	42,315	130.7	0.78
Gradient Boosting	39,762	124.8	0.80
Deep Learning (NN)	37,892	121.5	0.82

4.2 Applications of Medicaid Claims Forecasting

The superior performance of machine learning models, particularly random forests and neural networks, suggests that insurers can benefit from adopting these methods to predict claim amounts more accurately. However, the complexity of these models makes them less interpretable than traditional methods, which may be a barrier to adoption.

One potential solution is to use model-agnostic interpretation techniques, such as SHAP (Shapley Additive exPlanations), to improve the interpretability of machine learning models without sacrificing performance.

4.3 Interpretation and Practical Implications

Budgeting and Resource Allocation: Predicting future claims allows for better budgeting and allocation of resources to high-demand areas.

Healthcare Planning: Predictive models can inform healthcare providers and policymakers about potential surges in claims, helping to prepare for increased demand.

Fraud Detection: Machine learning can help identify suspicious claims, reducing Medicaid fraud and waste.

Policy Evaluation: By simulating the effect of policy changes, such as Medicaid expansion or changes in eligibility, predictive models can assess the impact on future claims.

5. Conclusion

This study demonstrates that machine learning algorithms can significantly improve the accuracy of health insurance claim prediction compared to traditional methods. While linear regression and decision trees offer simplicity and interpretability, more complex models such as random forests and neural networks provide higher predictive accuracy. Insurers should consider integrating machine learning techniques into their risk assessment and pricing strategies to remain competitive in a rapidly evolving market.

5.1 Future Work

Future work could focus on the integration of more advanced deep learning models, such as convolutional neural networks (CNNs) or long short-term memory networks (LSTMs), to capture temporal and spatial aspects of health data. Additionally, hybrid models that combine machine learning with traditional actuarial techniques could offer a balance between interpretability and performance.

1) 5.2 Challenges and Considerations

Data Quality: Medicaid data may contain missing, inaccurate, or delayed entries, which can negatively affect model performance.

Regulatory Constraints: The use of personal health information in ML models must comply with regulations like HIPAA (Health Insurance Portability and Accountability Act).

Interpretability: Some machine learning models, especially deep learning models, are considered "black boxes," making it difficult for policymakers to understand and trust the predictions.

Machine learning techniques offer a powerful toolset for forecasting Medicaid claims, enabling healthcare providers and policymakers to make data-driven decisions. By leveraging time series forecasting, regression models, and deep learning approaches, accurate predictions of Medicaid claim volumes and costs can be made, leading to more efficient resource allocation, better fraud detection, and improved overall management of Medicaid programs.

REFERENCES

[1] https://www.sciencedirect.com/science/article/pii/S2666827023000695

[2] https://www.ibm.com/topics/machine-learning

[3] https://www.tableau.com/learn/articles/top-machine-learning-methods

[4] https://www.mathworks.com/discovery/machine-learning.html

[5] Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems by Aurélien Géron