International Journal of Artificial Intelligence, Data Science, and Machine Learning
Grace Horizon Publication | Volume 2, Issue 1, 26-34, 2021
ISSN: 3050-9262 | https://doi.org/10.63282/3050-9262.IJAIDSML-V2I1P104

Original Article

An Advanced Machine Learning Models Design for Fraud Identification in Healthcare Insurance

Sriram Pabbineedi[1], Mitra Penmetsa[2], Jayakeshav Reddy Bhumireddy[3], Rajiv Chalasani[4], Mukund Sai Vikram Tyagadurgam[5], Venkataswamy Naidu Gangineni[6]
[1]University of Central Missouri.
[2]University of Illinois at Springfield.
[3]University of Houston.
[4]Sacred Heart University.
[5]University of Illinois at Springfield.
[6]University of Madras, Chennai.

Abstract - *Healthcare fraud threatens the interests of both healthcare facilities and the patients that they serve. In turn, such actions usually cause enormous financial damage and compromise the trustworthiness of healthcare systems. The proposed study intends, by adding a machine learning framework, to counter the issue of healthcare fraud. The systematic analysis of healthcare data shows patterns and anomalies which can be used to identify fraudulent behavior with more precision and speed. Based on the XG boost algorithm, this work describes a machine learning approach to irregularity detection and health insurance premium estimation. XG Boost algorithm was evaluated using the standardized performance measures such as R², MAE, RMSE and MAPE. Model's 89. 47% R² coupled with MAE of 1. XGBoost showed more predictive performance than Random Forest and Genetic Support Vector Machines (GSVMs) when compared. Support for generalization of the model was also offered when learning curves and prediction error plots were considered. From these results it is evident that XGBoost is a reliable approach for detecting insurance fraud and pricing within structured healthcare settings.*

Keywords - *Healthcare, insurance, Fraud Detection, Insurance fraud analytics, Artificial Intelligence, Insurance Claims, Machine Learning, Insurance fraud detection dataset.*

1. Introduction

The healthcare providers' capacity to provide quick medical services and work toward more efficient overall health performances is one of the key factors that support the promotion of society well-being. With healthcare services having to increase to accommodate growing populations, growing numbers of elderly people, and the continued escalation of chronic illnesses, the financial burden on both patients and healthcare providers continues to increase. Public and private insurance systems are important elements of the modern healthcare system as a result of attempts to reduce healthcare bills and make it possible for all to receive the necessary medical treatment. The purpose behind these systems is to provide all individuals with access to key treatments and services at an affordable price point while enabling providers to deliver such care effectively. However, as these benefits become of utmost urgency, healthcare insurance systems experience new vulnerabilities which include health insurance fraud emerging as a major concern [1][2]. Healthcare costs rise and the reliability of insurance programs is threatened by fraudulent claims, misrepresentations, and deliberate misuses of services [3]. Records show that a significant amount in billions is lost annually because of fraudulent activities that could support patients' healthcare needs [4]. The increasing incidence of health coverage deceit, inefficiency, and misuse are severely damaging the system's credibility and undermining its functional performance of the system.

Healthcare insurance fraud is complex and multifaceted, and very dynamic, which makes the fraud hard to detect and correct [5]. Policy loopholes and actions that bear a resemblance to real patient actions make it challenging for existing rule-based detection systems to maintain up to speed since the fraudsters adjust and take advantage of such openings [6][7]. Independent of their role being the groundwork for fraud detection, such systems tend to be lacking that flexibility and timeliness which is necessary to counteract the skillful methods used by today's fraud actors. Consequently, the healthcare insurance industry requires intelligent, scalable, and flexible solutions in order to identify and mitigate hidden places of fraud. Against this background, the use of complex methods of ML and DL as tools for healthcare insurance fraud detection has become a sound method. Using these technologies, it becomes possible to work with voluminous and complex data to dig out hidden trends, detect anomalies, and predict abnormal behaviors better and deeper. ML methods have proved able to distinguish between authentic and spurious claims, even within the complexity induced by forms of camouflage and occluded information [8][9]. The proposed framework uses conventional ML and

state-of-the-art DL methods for reducing false positives, improving detection precision, and dealing with modifications of fraud tactic. Incorporation of patient demographics, provider actions, claim profiles, and temporal forms enhances the model's relevance and makes its findings readily comprehensible.

1.1. Motivation and Contribution of Study

- In this study, a structured and scalable approach using XGBoost is presented for healthcare insurance premium estimation that can be easily adapted to detect fraudulent claims through anomaly detection.
- The study uses data downloaded from Kaggle which consists of a broad range of customer health and insurance characteristics, which facilitates meaningful analysis for premium prediction and fraud assignment.
- This study describes an all-inclusive pre-processing technique to data cleaning, ensuring data integrity, and applying the normalization Standard Scaler to ensure optimal input data for the model.
- Evaluating and ensuring EDA, vital associations such as the association between age, diabetes, and insurance expenses were identified, enhancing model clarity, and used to determine abnormalities or fraudulent data.
- Building upon the principal work by employing the XGBoost model, celebrated due to its accurate and rapid performance, this study proves its effectiveness within the framework of dealing with organized healthcare data in situations including fraud detection.
- The detailed analysis of the model based on R^2, MAE, RMSE, and MAPE confirmed its success and confirmed its usefulness in real-time fraud prediction systems.

1.2. Structure of Paper

The incorporation of this paper is as stated below. Section II discusses existing literature on the use of ML models in detecting fraud in healthcare insurance, Section III outlines the methodology as well as diagrams and techniques, Section IV explains the results and discusses them in context, and Section V concludes with results and suggestions for future directions on ML for fraud in healthcare insurance.

2. Literature Review

In this section, the research on health insurance fraud detection that makes use of ML. Rayan (2019) presents a hybrid system for identifying fraudulent claims from a given collection of outstanding claims that includes unsupervised learning (k-means clustering, outlier analysis), supervised learning (DT and Averaged Perceptron), and domain knowledge (Rule Engine). The investigation team is informed with a weighted priority queue of outstanding claims listing the most likely fraudulent claims with remarks for proactive and retrospective analysis. Their initial case study with one insurer demonstrates an increase in hit rate by 209.4% [1]. Johnson and Khoshgoftaar (2019) build automated fraud detection ML models. However, performance is hampered by issues with class-imbalanced huge data.

One of the techniques evaluated at the algorithm level is the Mean False Error Loss. Another is the Focal Loss. A cost-sensitive loss function is also used. Various class ratios are tested using different sample rates, and the optimal decision thresholds for each model are determined to achieve the required class-wise performance. Results of neural network tests on a 20% holdout dataset are explained by the area under the receiver operating characteristic curve (AUC). Overall, the findings show that ROS and ROS-RUS perform far better than baseline and algorithm-level methods, with AUC values of 0.8505 and 0.8509, respectively. ROS-RUS maximizes efficiency by reducing training time by a factor of four. It is found that decision boundaries are more stable when using algorithm-level techniques compared to baseline methods. Even basic RUS beats baseline methods with training time benefits of up to 30 times [10].

Sowah et al. (2019) identified health insurance fraud and other irregularities using the data pertaining to health insurance claims that were gathered from hospitals in Ghana. These health insurance databases employ GSVM, a new hybridized data mining and statistical ML technology that offers a suite of advanced algorithms for the automated identification of fraudulent claims. When used with SVM kernel classifiers, in terms of detection and classification, the experimental results showed that the GSVM performed better. The performance of three GSVM classifiers was assessed and contrasted. According to experimental results, using the different RBF kernel (87.91%), SVM classifiers (linear 80.67%), and polynomial 80.22 % significantly reduces the computing time required to process claims while improving classification accuracy [11].

Hung, Lin and Lee (2018) enhance stroke prediction in an extensive EMC database based on a population (552,898 patients), with a focus on the 25-45 age bracket. They use a new active data augmenter to develop a deep neural network model that predicts early strokes. The optimizer selects the most illuminating electronic health record samples from geriatric stroke patients. This strategy enhances the area under the receiver operating characteristic curve (AUC) value by 9.3% when compared to training directly using just young age group data and 8.2% when compared to training all age group data [12]. Hung et al. (2017) to forecast the occurrence of strokes within five years utilizing a large population-based EMC database of over 800,000 patients, compare DNN

with three other ML approaches. The results show that DNN and GBDT can achieve similar levels of accuracy in predictions as LR and SVM methods. In contrast, DNN uses less patient data than the GBDT approach while still producing the best outcomes [13].

Bauder and Khoshgoftaar (2017) compare a number of ML techniques to identify Medicare fraud. Four performance measures including the correction of class imbalance by oversampling and an 80-20 under-sampling technique are used in evaluation of mixed ML systems, as well as supervised and unsupervised ones. Classifying 2015 Medicare data according to provider categories is done using fraud labels from the List of Excluded Individuals/Entities database. Their research proves that detecting dishonest service providers is possible, students performed best when using the 80-20 sample method [14]. Table I provides a comparative overview of several ML approaches used for healthcare insurance fraud detection, which also highlights methodology, datasets, important discoveries, drawbacks, and recommendations for further study.

Table 1: Comparative Analysis of Machine Learning Techniques for fraud identification in healthcare insurance

Author	Methodology	Data	Key Findings	Limitation	Future Work
Rayan (2019)	k-means clustering, decision trees, averaged perceptrons, rule-based engines, and outlier analysis in a hybrid model	Claims dataset from one insurer	Achieved 209.4% increase in hit-rate; supports both proactive and retrospective fraud analysis	Limited to one insurer's data; generalizability unclear	Extend framework to multiple insurers and real-time fraud detection
Johnson and Khoshgoftaar (2019)	Comparison of 6 DL methods addressing class imbalance using ROS, RUS, ROS-RUS, and cost-sensitive loss functions	Medicare fraud detection dataset	ROS–RUS gave highest AUC (~0.8509); RUS reduced training time up to 30×; optimal thresholds strongly linked to class size	High class imbalance (minority class 0.03%); reliance on threshold tuning	Apply to other domains with class imbalance; improve threshold optimization strategies
Sowah et al. (2019)	Genetic SVMs using multiple kernel classifiers (linear, polynomial, RBF)	Ghana National Health Insurance Scheme dataset	RBF kernel achieved highest accuracy (87.91%); improved classification and reduced processing time	Limited to Ghana NHIS data; not benchmarked with deep learning models	Test GSVMs on larger, more diverse datasets; integrate with deep learning
Hung, Lin and Lee (2018)	Deep neural network with active data augmenter to select informative EHR samples	EMC database with 552,898 records	AUC improved by 9.3% and 8.2% over baseline training on young-only and all-age data respectively	Focused on stroke prediction, not fraud; model domain-specific	Apply active augmentation to fraud detection in low-sample settings
Hung et al. (2017)	Comparison of DNN, GBDT, LR, and SVM for stroke prediction	EMC database with 800,000 patients	DNN and GBDT performed best; DNN required less data for optimal results	Health outcome prediction domain; not focused on fraud	Extend model comparisons to insurance fraud datasets
Bauder and Khoshgoftaar (2017)	Supervised, unsupervised, and hybrid ML models; class imbalance handled via oversampling and 80-20 undersampling	2015 Medicare data with labels from Excluded Individuals/Entities	80-20 undersampling method yielded best fraud detection performance	Class imbalance still an issue; potential underutilization of unsupervised methods	Enhance hybrid models; explore newer imbalance-handling algorithms

3. Methodology

This study adopts a systematic approach for the identification of healthcare insurance fraud by using ML algorithms, specifically targeting the XGBoost model. The data utilized for this study was accessed from Kaggle and included 986 records with 11 attributes of customer health and insurance information. It was clear from doing EDA that there were individual relationships between age and diabetes and the premium paid. The pre-processing phase confirmed the lack of missing or duplicate entries and endorsed the overall

quality of the dataset. Feature scaling with Stander Scaler. There were 25% testing and 75% training subsets in the dataset. The XGBoost algorithm, known for its scalability and performance, was applied to predict insurance premiums based on health-related features. Model evaluation was conducted using to evaluate the correctness of the model, use R2, MAE, RMSE, and MAPE measures, interpretability, and prediction quality. This whole process is shown in Figure 1:

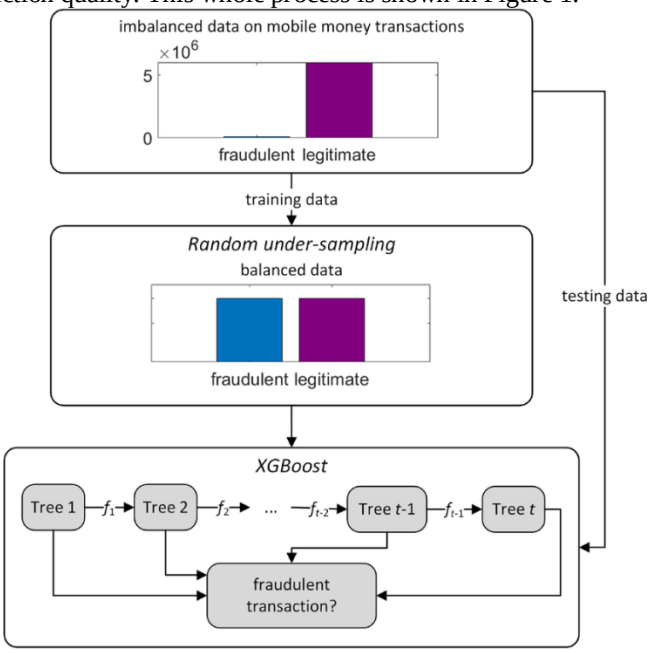

Fig 1: Flowchart of Fraud Identification in Healthcare Insurance

The following sections provide each step description which also shows in methodology and proposed flowchart:

3.1. Data Collection

The KAGGLE repository served as the source of the medical insurance cost dataset. There are 11 attributes/features and 986 records in the collection. EDA was used to swiftly examine the information in an effort to find any hidden patterns, identify anomalies, test theories, and verify presumptions. The Pearson correlation heatmap in Figure 2 illustrates how certain important characteristics are related to one another in order to ascertain their association.

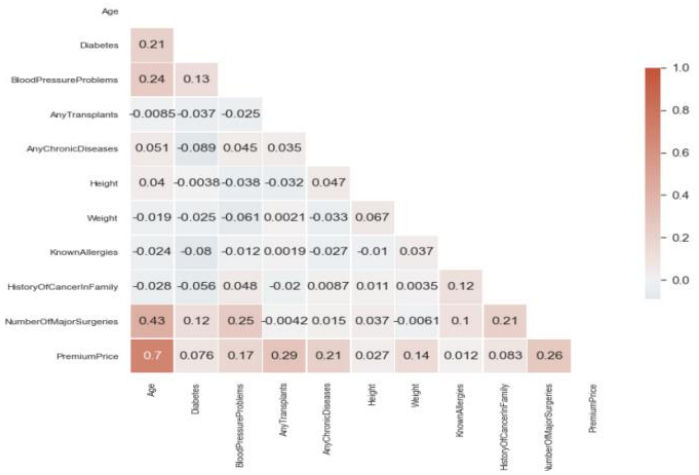

Fig 2: Correlation Heatmap

Figure 2 presents a correlation heatmap depicting the relationships among various health-related and demographic variables, Age, blood pressure issues, diabetes, any transplants, any chronic illnesses, height, weight, known allergies, a family history of cancer, the number of major surgeries, and the premium price are all included. Stronger correlations are shown by more vivid colors in the heatmap, which employs a color scale ranging from blue (strongly correlated negatively) to red (strongly correlated positively). Notably, Premium Price shows a moderate positive correlation with Age (0.70), Blood Pressure Problems (0.26), and

Diabetes (0.27), indicating that these factors may significantly influence premium pricing. Other correlations are generally weak, suggesting limited linear relationships between most variable pairs.

Fig 3: Insurance Premium Price by Age

Figure 3 illustrates the relationship between age and insurance premium price, showing a clear upward trend as age increases. The plot indicates that younger individuals, particularly those under 30, tend to pay lower premiums, while a noticeable rise in premium prices occurs after age 30. This increase continues steadily through middle and older age groups, with premiums reaching their highest levels beyond age 60. The shaded region around the line represents the confidence interval, suggesting variability in premium prices within each age group but reaffirming the overall positive correlation between age and insurance premium cost.

3.2. Data Preprocessing

Preprocessing data is seen as an essential stage in both data mining and ML [15]. Data that is noisy, lacking, inconsistent, or redundant is frequently found in large datasets. To create a reliable model, the data must go through a number of pre-processing steps to get it into the right format.
These are the key pre-processing words:

3.3. Check Missing and Duplicate Value

To find any missing or duplicate entries, a first evaluation of the dataset was conducted. The dataset was carefully examined for data quality issues, and it was confirmed that there were no duplicate records present. This ensured the reliability and consistency of the dataset for further analysis.

3.4. Feature Scaling with Stander Scaler

Feature scaling using Standard Scaler was applied to normalize the input data before training the model. Standard Scaler creates a normal distribution where the mean is zero and the standard deviation is one, achieved by homogenizing the variables by removing the mean and scaling them to the unit variance. One popular technique for standardizing data is Z-score normalization, which is as follows Equation (1).

$$z = \frac{x-\mu}{\sigma} \qquad (1)$$

The feature mean and standard deviation are represented by μ and σ, respectively, whereas x represents the original value and z represents the standardised value.

3.5. Data Splitting

The training and test datasets were then created by dividing the data into two halves. About 25% of the entire data is utilized for testing, while the remaining 75% is used for training.

3.6. Proposed XGBoost

The gradient boosting framework, which was first created, was converted into an effective and scalable implementation with the XGBoost model. Through the use of boosting, XGBoost seeks to increase model performance and computing speed [16]. Given a dataset with n samples and m characteristics, D $=\{(x_{i,y_i})|\, x_i \in R^m, y_i \in R\}$ and $\{x_{i1}, x_{i2}, \ldots, x_{im} \,|i = 1,22, \ldots, n\}$. XGBoost constructs t trees in a way that it predicts value $\hat{y}_i^{(t)}$ up to According to Equations (2–5), the t-th tree [17].

$$\hat{y}_i^{(0)} = 0 \qquad (2)$$

$$\hat{y}_i^{(1)} = f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \qquad (3)$$

$$\hat{y}_i^{(2)} = f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i) \quad (4)$$
$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} = f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \quad (5)$$

The approach generates a weak classifier $f_k(x_i) =$ for each iteration, and the predicted value $\hat{y}_i^{(t)}$ of each iteration is the sum of the decision tree result of this round $\hat{y}_i^{(t-1)}$ and the anticipated worth of the prior cycle $f_t(x_i)$. Additionally, unlike basic gradient boosting techniques [18], One weak learner at a time is not added by XGBoost. Rather, the method uses a multi-threaded approach, which makes sure the CPU core of the system is used appropriately and, consequently, increases speed and overall performance.

3.7. Performance Evaluation Metrics

An essential part of building ML projects is evaluating the models, which helps to comprehend how well they work and makes it easier to explain and show the results. Since the exact value of a regression model is not always easy to predict, the goal is to show how well the predicted values correspond to the real values. In this study, the models were evaluated using four performance assessment criteria: R2, MAE, RMSE, and MAPE. These metrics determine as:

R-Square: R^2, or the coefficient of determination, shows what proportion of the variance in the dependent variable can be explained by the independent variables. It is a model fit metric, Equation (6).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} where, \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad (6)$$

Mean Absolute Error (MAE): Evaluates forecast errors' average magnitude without taking direction into account. It is computed by averaging the absolute disparities between the values that were anticipated and those that were real. It's a measure of model fit Equation (7):

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |(y_i - y_i^p)| \quad (7)$$

Root Mean Squared Error (*RMS*E): In order to return the metric to RMSE, which is the square root of MSE, is expressed in the same units as the target variable. It offers a measure of average error size that is simple to understand. It's a measure of model fit Equation (8).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^p)^2} \quad (8)$$

Mean Absolute Percentage Error (MAPE): A metric for accuracy based on percentages, MAPE calculates the average percentage error between actual and anticipated values. An effective method for evaluating relative error rates, which indicate how well the model works, is MAPE. It's a measure of model fit Equation (9).

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n} |(\frac{y_i - y_p^i}{y_i})| * 100 \quad (9)$$

Assess the suggested XGBoost model's health insurance performance, and think about how these improvements may be made.

4. Results And Discussion

The results of using the suggested XGBoost model for health insurance are presented in the Table II. The system included a GPU, 32 GB of RAM, an Intel Core i7-13900K CPU, and Windows 11 Pro as its operating system. Below, the suggested model's outcomes are examined in relation to performance metrics and contrasted with those of other models. 89.470% R^2 indicates that the identified model explains a large portion of the variation of the target variable. The MAE of 134.80 and RMSE of 22.320 reflect relatively low average prediction errors, with RMSE highlighting the model's robustness against larger errors. Moreover, a MAPE of 4.608% shows that the model perfectly delivers high accuracy for insurance claims of different sizes. With these results combined, it therefore is evident that that XGBoost is well adapted and strong enough to produce good results in insurance predictions.

Table 2: Experiment Results of XGBoost Model for Health Insurance Prediction

Measures	XGBoost
R^2	89.470
MAE	134.80
RMSE	22.320
MAPE	4.608

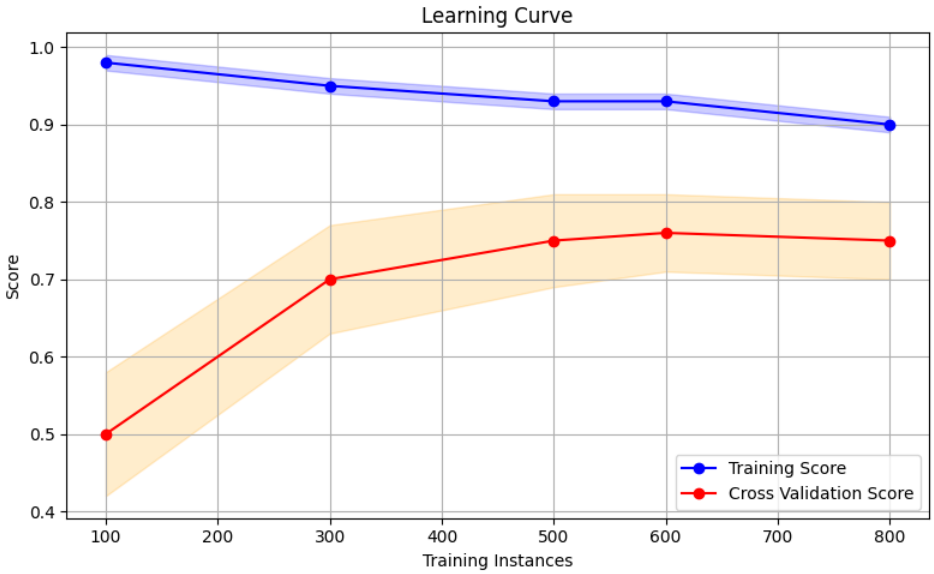

Fig 4: Plot Learning Curve for XGB

Figure 4 displays the learning curve for the XGBoost (XGB) model, illustrating how model performance varies with the number of training instances. The training score (in blue) starts near 1.0 with a small dataset and gradually declines as the number of training instances increases, indicating reduced overfitting. Meanwhile, the cross-validation score (in red) improves significantly with more training data, rising from around 0.5 to approximately 0.75, and then stabilizes. The narrowing gap between the training and cross-validation scores as data increases suggests improved generalization of the model with larger training sets. Shaded regions indicate confidence intervals around the scores.

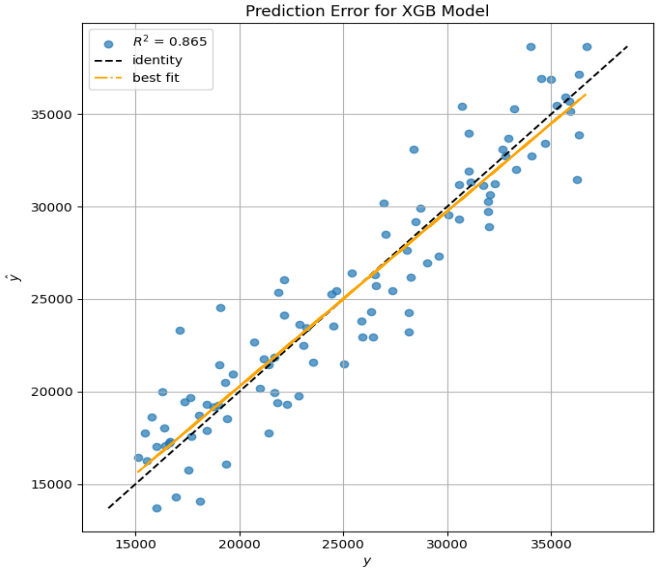

Fig 5: Prediction Error Plot for XGBoost Model

Figure 5 shows a prediction error plot for the XGBoost model displays the relationship between the actual values (y-axis) and the predicted values (y-axis) in relation to one another. The scattered blue points represent individual predictions, while the dashed black 'identity' line indicates perfect predictions. The orange 'best fit' line represents the linear regression fit through the predicted values, and its proximity to the identity line, along with the high R^2 value of 0.865, suggests a strong positive correlation and good predictive performance of the XGBoost model, although some degree of prediction error is still evident in the scatter of points around the lines.

Table 3: ML Models Comparison for Health Insurance Prediction

Measures	**R^2**
XGBoost	89.470
Random Forest[19]	0.87
Genetic Support Vector Machines (GSVMs)[11]	87.91

In terms of predictive accuracy, Table III's performance comparison of ML models for health insurance prediction shows that the XGBoost model performs better than both RF and GSVM. XGBoost achieved the highest R^2 score of 89.47%, indicating its superior ability to explain the variance in insurance premium predictions. In comparison, RF attained an R^2 of 0.87, while GSVMs reported a slightly lower R^2 of 87.91%. The results show that XGBoost is suitable for structured healthcare data and provides greater prediction accuracy, making it a strong candidate for insurance fraud detection and premium estimation tasks.

5. Conclusion And Future Scope

In recent years, there have been massive financial losses on the part of the medical insurance companies due to fraudulent activities. By integrating these firms' fraud detection procedures with ML, it is likely to have better outcomes. This work validates the effectiveness of XGBoost in both cost estimation of healthcare insurance and identification of fraudulent patterns in structured medical data sets. To achieve this, a rigorous workflow needed to be put in place that includes data preprocessing and exploratory analysis and a careful model evaluation of the XGBoost algorithm which delivered high accuracy with an R^2 of 89.47%, MAE of 134.80, RMSE of 2 These findings highlight the high parameter and the ability of the model to sustain performance for diverse data patterns. Overall, overall, the XGBoost algorithm provided more accurate and reliable results than those observed in RF and Genetic SVM, and other ML algorithms. These learning and prediction error analyses provided the backing of the model's robustness and avoided overfitting control problems. Because of the small size of the dataset, as well as the exclusive usage of structured features, this study's shortfalls can impact the scalability and transferability of the model to other contexts more broadly. The next phase of research focuses on the use of comprehensive, current datasets in combination with natural language processing in order to improve the fraud detection process and increase overall predictive results.

References

[1] N. Rayan, "Framework for analysis and detection of fraud in health insurance," in *Proceedings of 2019 6th IEEE International Conference on Cloud Computing and Intelligence Systems, CCIS 2019*, 2019. doi: 10.1109/CCIS48116.2019.9073700.

[2] S. Chen and A. Gangopadhyay, "A novel approach to uncover health care frauds through spectral analysis," in *Proceedings - 2013 IEEE International Conference on Healthcare Informatics, ICHI 2013*, 2013. doi: 10.1109/ICHI.2013.77.

[3] S. Kareem, R. B. Ahmad, and A. B. Sarlan, "Framework for the identification of fraudulent health insurance claims using association rule mining," in *2017 IEEE Conference on Big Data and Analytics, ICBDA 2017*, 2017. doi: 10.1109/ICBDAA.2017.8284114.

[4] V. Rawte and G. Anuradha, "Fraud detection in health insurance using data mining techniques," in *Proceedings - 2015 International Conference on Communication, Information and Computing Technology, ICCICT 2015*, 2015. doi: 10.1109/ICCICT.2015.7045689.

[5] C. Sun, Q. Li, H. Li, Y. Shi, S. Zhang, and W. Guo, "Patient Cluster Divergence Based Healthcare Insurance Fraudster Detection," *IEEE Access*, vol. 7, pp. 14162–14170, 2019, doi: 10.1109/ACCESS.2018.2886680.

[6] A. Bayerstadler, L. van Dijk, and F. Winter, "Bayesian Multinomial Latent Variable Modeling for Fraud and Abuse Detection in Health Insurance," *Insur. Math. Econ.*, vol. 71, pp. 244–252, Nov. 2016, doi: 10.1016/j.insmatheco.2016.09.013.

[7] A. Verma, A. Taneja, and A. Arora, "Fraud detection and frequent pattern matching in insurance claims using data mining techniques," in *2017 10th International Conference on Contemporary Computing, IC3 2017*, 2017. doi: 10.1109/IC3.2017.8284299.

[8] P. Doupe, J. Faghmous, and S. Basu, "Machine Learning for Health Services Researchers," *Value Heal.*, 2019, doi: 10.1016/j.jval.2019.02.012.

[9] C. Francis, N. Pepper, and H. Strong, "Using support vector machines to detect medical fraud and abuse," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2011. doi: 10.1109/IEMBS.2011.6092044.

[10] J. M. Johnson and T. M. Khoshgoftaar, "Medicare fraud detection using neural networks," *J. Big Data*, 2019, doi: 10.1186/s40537-019-0225-0.

[11] R. A. Sowah *et al.*, "Decision Support System (DSS) for Fraud Detection in Health Insurance Claims Using Genetic Support Vector Machines (GSVMs)," *J. Eng. (United Kingdom)*, 2019, doi: 10.1155/2019/1432597.

[12] C. Y. Hung, C. H. Lin, and C. C. Lee, "Improving Young Stroke Prediction by Learning with Active Data Augmenter in a Large-Scale Electronic Medical Claims Database," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2018. doi: 10.1109/EMBC.2018.8513479.

[13] C. Y. Hung, W. C. Chen, P. T. Lai, C. H. Lin, and C. C. Lee, "Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2017. doi: 10.1109/EMBC.2017.8037515.

[14] R. A. Bauder and T. M. Khoshgoftaar, "Medicare Fraud Detection Using Machine Learning Methods," in *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2017, pp. 858–865. doi: 10.1109/ICMLA.2017.00-48.

[15] A. Immadisetty, "Edge Analytics vs. Cloud Analytics: Tradeoffs in Real-Time Data Processing," *J. Recent Trends Comput. Sci. Eng.*, vol. 13, no. 1, pp. 42–52, 2016.

[16] A. H. Anju, "Extreme Gradient Boosting using Squared Logistics Loss function," *Int. J. Sci. Dev. Res.*, vol. 2, no. 8, pp. 54–61, 2017.

[17] L. Torlay, M. Perrone-Bertolotti, E. Thomas, and M. Baciu, "Machine learning–XGBoost analysis of language networks to classify patients with epilepsy," *Brain Informatics*, 2017, doi: 10.1007/s40708-017-0065-7.

[18] R. Tarafdar and Y. Han, "Finding Majority for Integer Elements," *J. Comput. Sci. Coll.*, vol. 33, no. 5, pp. 187–191, 2018.

[19] S. Suri and D. V Jose, "Effective Fraud Detection in Healthcare Domain using Popular Classification Modeling Techniques," *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 11, 2019.

[20] Kalla, D., & Samiuddin, V. (2020). Chatbot for medical treatment using NLTK Lib. *IOSR J. Comput. Eng*, *22*, 12.

[21] Kuraku, S., & Kalla, D. (2020). Emotet malware a banking credentials stealer. *Iosr J. Comput. Eng*, *22*, 31-41.