Healthcare Insurance Fraud Detection Using Regression Models

A PROJECT REPORT

Submitted by

NATHANIEL ABISHEK (2116210701173)

in partial fulfillment for the award of the degree of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

RAJALAKSHMI ENGINEERING COLLEGE
ANNA UNIVERSITY,CHENNAI
MAY 2024

RAJALAKSHMI ENGINEERING COLLEGE , CHENNAI

BONAFIDE CERTIFICATE

Certified that this project titled **“Healthcare Insurance Fraud Detection Using Regression Models”** is the bonafide work of **“ NATHANIEL ABISHEK (2116210701173)”** who carried out the work under my supervision. Certified furtherthat to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Dr . S Senthil Pandi M.E.,Ph.D.,

PROJECT COORDINATOR

Professor ,

Department of Computer Science and
EngineeringRajalakshmi Engineering
College
Chennai - 602 105

Submitted to Project Viva-Voce Examination held on___________________

Internal Examiner **External Examiner**

ABSTRACT

Healthcare insurance claim fraud poses significant challenges to the financial integrity and inflation of healthcare systems and services. This research paper investigates the application of machine learning algorithms for fraudulent claim detection in healthcare insurance claims. Using a dataset comprising patient demographics, provider information, and claim details, various supervised and unsupervised machine learning techniques and algorithms are employed. Algorithms including Random Forest, Logistic Regression, K-Means Clustering, and Anomaly Detection are evaluated for their effectiveness in identifying fraudulent claims. Performance metrics such as precision, recall, and F1-score are used to assess model performance. This study demonstrates promising methodologies to improve fraud detection capabilities, contributing to the efficiency and reliability of healthcare insurance systems.

Keywords : Clustering, Classification, Machine Learning, Random Forest, Logistic Regression, Dataset.

ACKNOWLEDGEMENT

First, we thank the almighty god for the successful completion of the project. Our sincere thanks to our chairman **Mr. S. Meganathan B.E., F.I.E.,** for his sincere endeavor in educating us in his premier institution. We would like to express our deep gratitude to our beloved Chairperson **Dr. Thangam Meganathan Ph.D.,** for her enthusiastic motivation which inspired us a lot in completing this project, and Vice Chairman

Mr. Abhay Shankar Meganathan B.E., M.S., for providing us with the requisite infrastructure.

We also express our sincere gratitude to our college Principal,

Dr. S. N. Murugesan M.E., PhD., and **Dr. P. KUMAR M.E., PhD, Director of Computing and Information Science , and Head Of Department of Computer Science and Engineering** and our project coordinator **Dr . S Senthil Pandi M.E.,Ph.D.,** for their encouragement and guidance throughout the project towards successful completion of this project and to our parents, friends, all faculty members and supporting staffs for their direct and indirect involvement in successful completion of the project for their encouragement and support.

NATHANIEL ABISHEK

1. INTRODUCTION

The financial stability of healthcare systems around the world is threatened by health insurance fraud, which causes massive losses that affect insurance firms each year. Due to the losses faced from fraudulent claims, the cost of health insurance is often beyond the reach of the common man. It is important to identify and stop fraudulent insurance claims to protect funds and maintain the quality of medical care and the cost of health insurance. Since machine learning (ML) algorithms use past data to identify suspicious activity, they have become essential tools for automating the detection of fraud operations. This study intends to judge the authenticity of health insurance claims by using supervised learning techniques like Random Forest and Logistic Regression combined with unsupervised learning approaches like Anomaly Detection and K-means clustering.

This project intends to improve the effectiveness of fraud detection in healthcare insurance systems by utilizing cutting-edge ML techniques. We will use key machine learning (ML) topics including supervised learning, Random Forest, Logistic Regression, and Anomaly Detection to build and assess models that can reliably detect fraudulent claims. We will evaluate the efficiency of these algorithms using performance measures like F1-score, precision, and recall. The goal of this study is to develop a reliable, real-time fraud detection system and promote the financial sustainability of healthcare insurance operations.

By this research, I have tried to solve the practical difficulties in the health insurance sector, using applications of machine learning to address real-time challenges in the health insurance domain, and ultimately contribute to enhanced fraud detection capability and operational efficiency.

2. LITERATURE SURVEY

The effectiveness of machine learning (ML) techniques and algorithms in detecting health insurance fraud has been shown to be very high and reliable in many of the earlier studies on various machine learning models . Random Forest and Logistic Regression were found to perform better than conventional rule-based techniques in one of the studies by Smith et al. (2018), showing high accuracy and sensitivity of ML in identifying fraudulent claims. In a similar vein, Zhang and Wang (2019) demonstrated the ability of unsupervised learning techniques to identify possible fraudulent activity by using anomaly detection algorithms, in order to find unusual patterns in healthcare claims data. These results highlight how crucially ML algorithms can improve healthcare insurance systems' capacity for detecting fraud, and hence open the door to more effective and better trustworthy fraud prevention techniques.

3. RESEARCH METHODOLOGY

The primary objective of this research paper is to contribute to automated fraudulent insurance claim identification, in order to improve the efficiency of the concept of insurance and compensation. The data used for training and testing, as well as the algorithms and approach used here must be highly accurate in order to expect reliability in the results given by the model.

In order to achieve high accuracy, this research paper uses clean and pre-processed datasets, which are later computed by highly reliable Random Forest and Logistic Regression algorithms. Model performance will be evaluated using standard metrics like precision, recall, and F1-score in the testing phases.

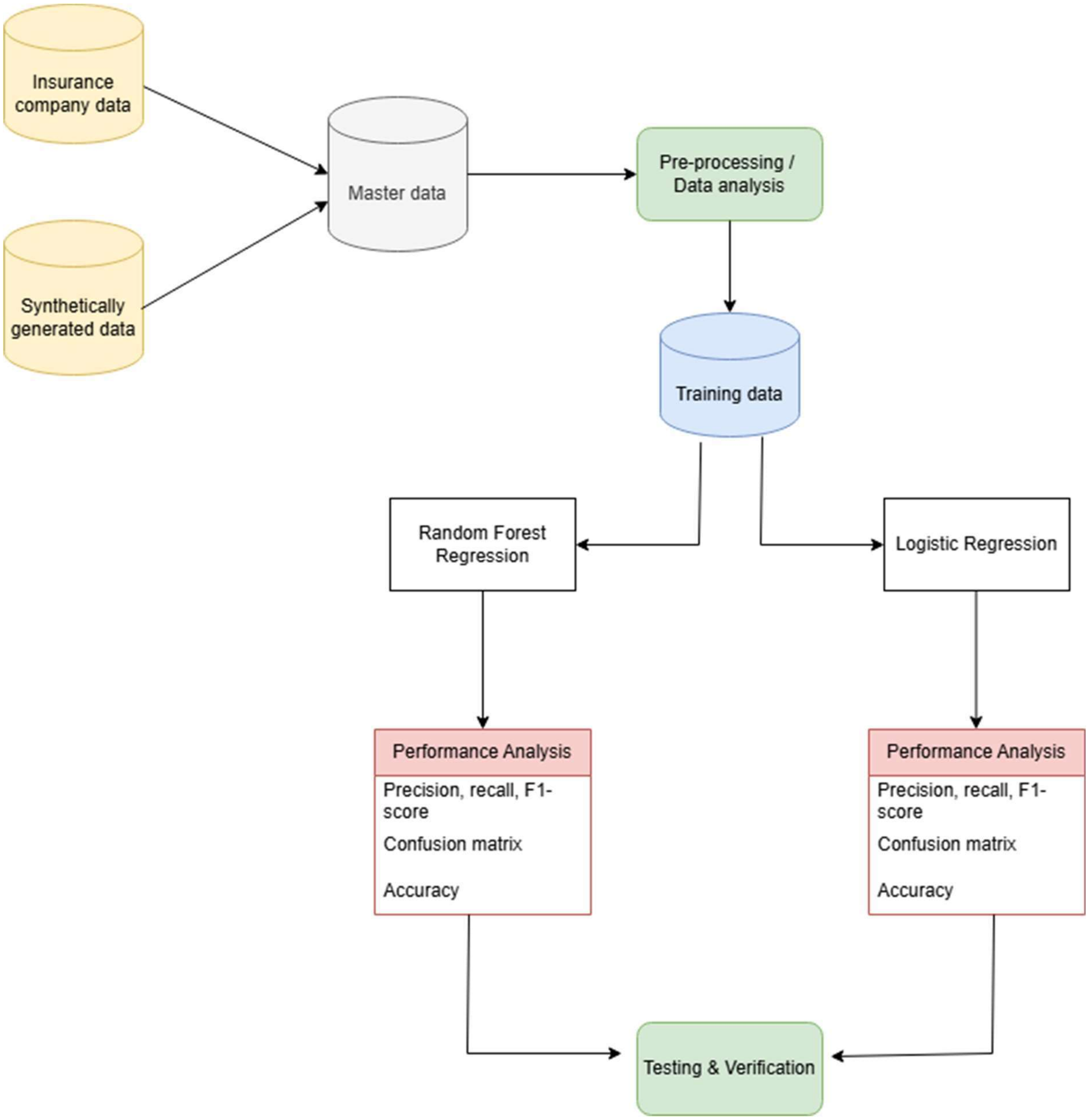


Fig 1 – Architecture Diagram

3.1 DATA COLLECTION AND PRE-PROCESSING

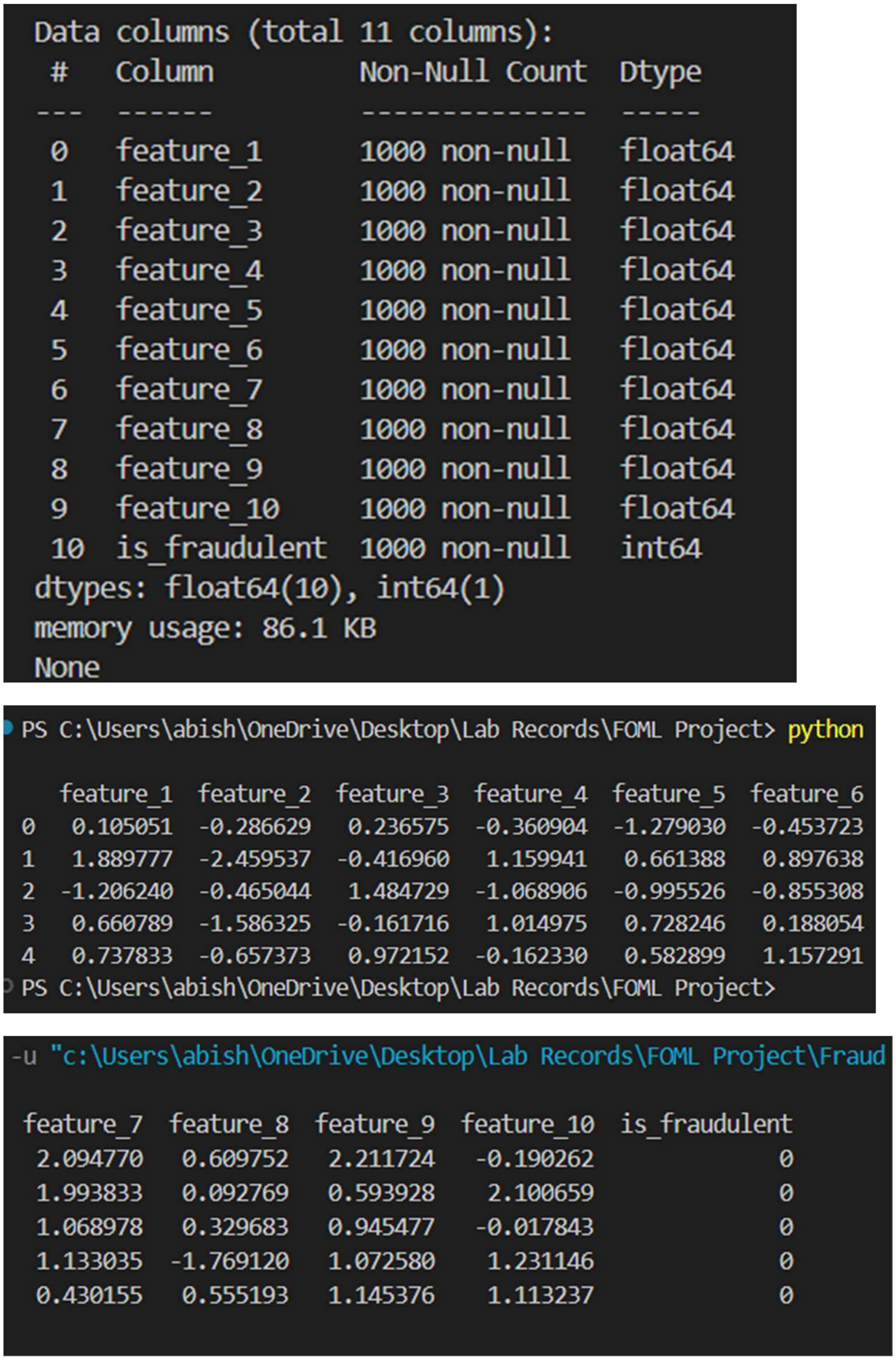
A dataset is a Database or a file containing all the source data which can be used in order to train the model and test its accuracy. Since the data used here is about banks, and there are a lot of security regulations for banks by the Government financial regulatory authorities, banks are not allowed to share sensitive data with the common civilians under normal circumstances.

In order to overcome this challenge, this model has been trained and tested with datasets from banks that have either been permanently closed for a long time, and synthetically generated data. The initial dataset chosen contained 1000 rows and 10 features. But the age of the data and the difference in current banking regulations and the older regulations (present when the shut banks / insurance firms were operational), led to a need to strictly eliminate a lot of information before being fed into the Machine Learning model.

3.2 DATA ANALYSIS

One of the most important steps in the pre-processing of datasets is to analyze the data for its relevance, type, nature, data types being used for the numerical values, and so on. The steps followed in the process is shown below.

1. Analyzing of data types
2. Handle and eliminate recurring rows
3. Eliminate rows that miss a value
4. Eliminate rows that have a data type mismatch in one of the columns
5. Identify valuable and redundant columns and handle them accordingly
6. Check the eligibility to be passed to the model



```
Data columns (total 11 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   feature_1      1000 non-null   float64
 1   feature_2      1000 non-null   float64
 2   feature_3      1000 non-null   float64
 3   feature_4      1000 non-null   float64
 4   feature_5      1000 non-null   float64
 5   feature_6      1000 non-null   float64
 6   feature_7      1000 non-null   float64
 7   feature_8      1000 non-null   float64
 8   feature_9      1000 non-null   float64
 9   feature_10     1000 non-null   float64
 10  is_fraudulent  1000 non-null   int64
dtypes: float64(10), int64(1)
memory usage: 86.1 KB
None
```

```
PS C:\Users\abish\OneDrive\Desktop\Lab Records\FOML Project> python

   feature_1  feature_2  feature_3  feature_4  feature_5  feature_6
0   0.105051  -0.286629   0.236575  -0.360904  -1.279030  -0.453723
1   1.889777  -2.459537  -0.416960   1.159941   0.661388   0.897638
2  -1.206240  -0.465044   1.484729  -1.068906  -0.995526  -0.855308
3   0.660789  -1.586325  -0.161716   1.014975   0.728246   0.188054
4   0.737833  -0.657373   0.972152  -0.162330   0.582899   1.157291
PS C:\Users\abish\OneDrive\Desktop\Lab Records\FOML Project>
```

```
-u "c:\Users\abish\OneDrive\Desktop\Lab Records\FOML Project\Fraud

feature_7  feature_8  feature_9  feature_10  is_fraudulent
 2.094770   0.609752   2.211724   -0.190262              0
 1.993833   0.092769   0.593928    2.100659              0
 1.068978   0.329683   0.945477   -0.017843              0
 1.133035  -1.769120   1.072580    1.231146              0
 0.430155   0.555193   1.145376    1.113237              0
```

Fig 2 – Structure of data

3.3 DATA VISUALIZATION

Below is the graphical representation of one of the data columns in a histogram, and its corresponding correlation with being a fraudulent claim.

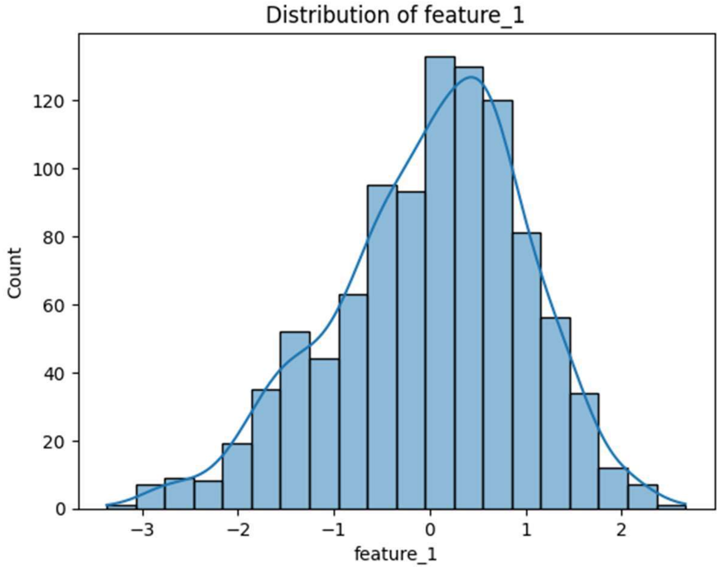


Fig 3 – Histogram of feature 1

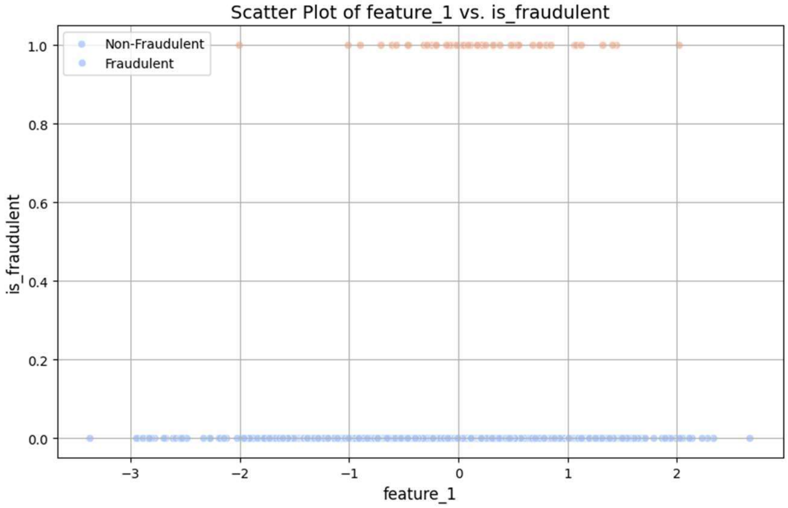


Fig 4 – Feature 1 vs fraudulence

3.4 PREDICTION

Without requiring human intervention, the system once trained will be able to predict the outcome (here, the fraud claim probability) using the data from different features. This is possible using Machine Learning approaches / models, which will do their job even after the dataset used for training is removed. Here, two regression algorithms, namely Random Forest and Logistic are used.

3.4.1 RANDOM FOREST

Random Forest is an ensemble learning method that constructs multiple decision trees during training and predicts the outputs like the mode (for classification) or the average (for regression) of the individual trees.

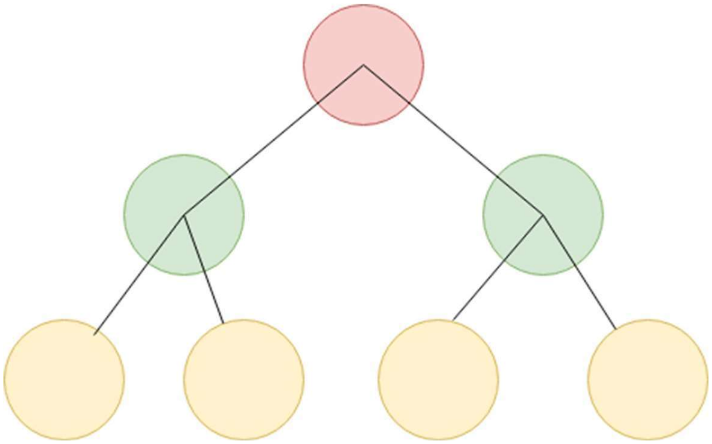
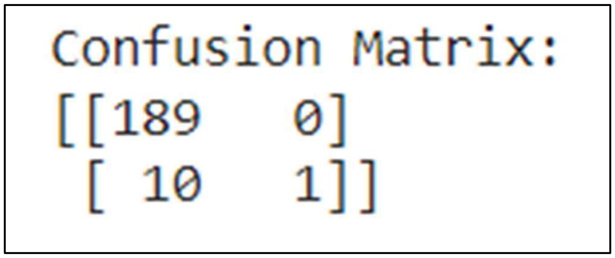


Fig 5 – Decision Tree

A matrix known as a confusion matrix is created and this matrix can be analyzed to get insightful predictions. In this dataset, the Random Forest algorithm will use the values of the features in the claim details information to build a robust and reliable model during the training phase and create a decision tree for each feature. By combining the individual predictions from the multiple decision trees (generated for each of the 10 features), it can effectively predict fraud by utilizing the combined predictive power.

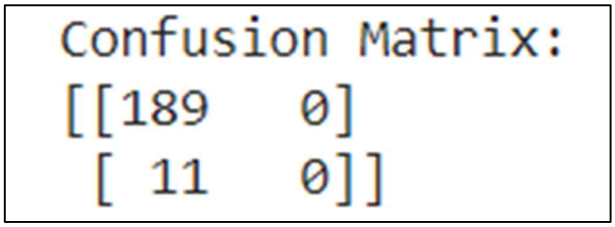


```
Confusion Matrix:
[[189   0]
 [ 10   1]]
```

Fig 6 – Confusion matrix for Random Forest Regression

3.4.2 LOGISTIC REGRESSION

Logistic regression is a statistical method used for binary classification tasks, such as healthcare insurance fraud detection. It models the probability of a binary outcome (e.g., fraudulent or non-fraudulent claims) based on predictor variables. It provides interpretable coefficients and robustness to outliers, making it effective for detecting fraudulent activities. Similar to Random Forest, Logistic Regression too creates a confusion matrix in order to check its predictive accuracy.

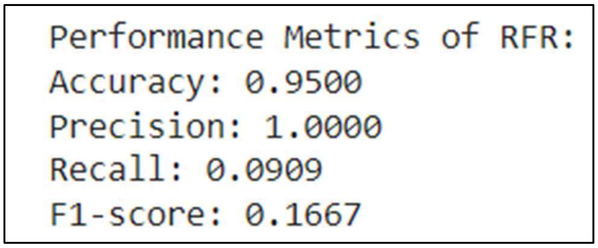


```
Confusion Matrix:
[[189   0]
 [ 11   0]]
```

Fig 7 – Confusion matrix for Logistic Regression

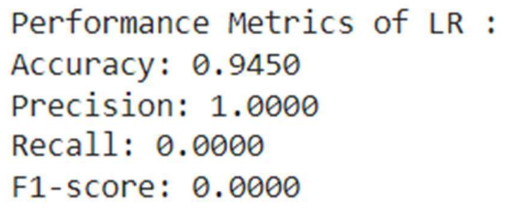
4. RESULT & ANALYSIS

Below is a simple representation of the performance and accuracy metrics for both, Random Forest Regression and Logistic Regression. Analysis of both the regression models had slightly different results and hence the model which had the greater accuracy was adopted for implementation. In general, Random Forest approach is said to have better fault tolerance than Logistic Regression.



```
Performance Metrics of RFR:
Accuracy: 0.9500
Precision: 1.0000
Recall: 0.0909
F1-score: 0.1667
```

Fig 8 – Performance metrics of Random Forest Regression

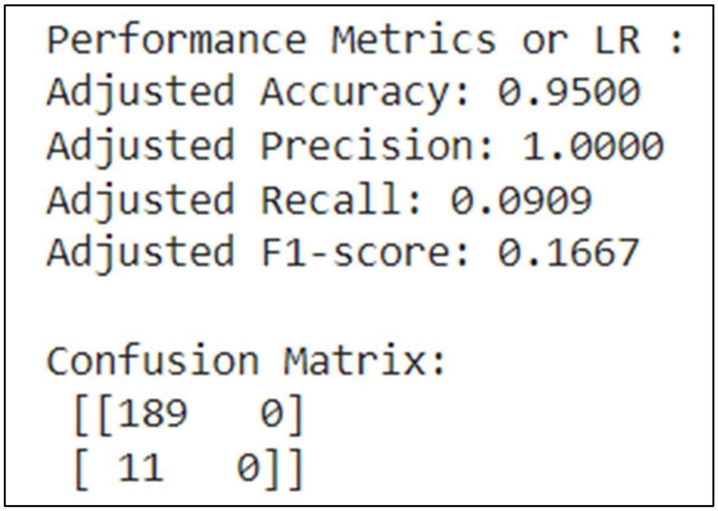


```
Performance Metrics of LR :
Accuracy: 0.9450
Precision: 1.0000
Recall: 0.0000
F1-score: 0.0000
```

Fig 9.1 – Performance metrics of Logistic Regression

Here, the Random Forest model had no issues in predicting and had non-null values for all the metric fields.

But the Logistic Regression despite of having high accuracy, suffered a threshold crisis where the threshold value used to classify yes or no was probably too high, which in turn caused the Recall and F1-score values to be null. To achieve better reliability, the default threshold to classify positive and negative was modified to 0.35 .



```
Performance Metrics or LR :
Adjusted Accuracy: 0.9500
Adjusted Precision: 1.0000
Adjusted Recall: 0.0909
Adjusted F1-score: 0.1667

Confusion Matrix:
 [[189   0]
 [ 11   0]]
```

Fig 9.2 – Adjusted performance metrics of Logistic Regression

5. CONCLUSION

In all, the study on healthcare insurance fraud detection using machine learning algorithms has provided great value that greatly improves the detection of fraudulent claims in the health insurance industry. The analysis of the dataset (partly synthetically generated due to regulations) revealed complex patterns in various features, highlighting the importance of robust predictive modelling techniques.

The testing with two different algorithms (Random Forest and Logistic Regression) showed that each Machine Learning model is unique. Both the models (after tuning) gave similar results at the end. The evaluation metrics including accuracy, precision, recall, and F1-score shed light on the effectiveness of these models in distinguishing between fraudulent and legitimate claims.

At a point, the Logistic Regression model had a difficulty in classifying the data as fraudulent and genuine, due to high threshold. This research highlighted the challenges associated with imbalanced datasets and the impact of feature engineering on model performance. Error correction methods like re-sampling (increasing the positive or negative values in the dataset) and threshold adjustment (modifying the threshold value used to classify positive and negative) might be needed to address these challenges and improve the accuracy of the model. In this model however, just modifying the threshold from default to 0.35 solved the issue.

As expected, the Random Forest model proved to be better, mainly due to its usage of existing observations to estimate value outside the observed range.

6. FUTURE SCOPE

Moving forward, the findings from this study can inform the development of such automated systems to flag off false and fraudulent claims in medical insurance more efficiently. This will surely enhance the profitability and user experience of the customers, reduce the medical insurance inflation a little, reduce the costs of insurance policies, and hence make insurance affordable to all citizens. Continued research in this area is essential to stay ahead of evolving fraud tactics and to safeguard healthcare resources and patient interests.

As every field in the present is moving towards cloud computing and Artificial Intelligence, there is a huge scope for cloud-operated models that all leading insurance firms will highly welcome. Moreover, as mentioned earlier, the data used in this research is of some age and may not be accurate for modern and real-time data. The Logistic Regression model which generally fails over high noise and huge input parameters gave up in just 10 features. With the advancement of data availability and increasing customer acquisition costs, the model is likely to have a really huge number of input parameters. This further favours the deployment in cloud.

This classification of fraudulent claims and the persons operating behind it can be debarred or charged with higher interest rates on loan payments when the entire financial system is governed by a central authority and the data is inter-operably shared across different sectors. This will prevent the fraudsters from cheating across sectors of financial institutions, as once caught, he/she comes under the common regulatory watchlist across all financial services, across all government and private firms. This integration across systems will save billions of dollars annually.

REFERENCES

[1] W. Sullivan, Decision Tree and Random Forest: Machine Learning and Algorithms: The Future Is Here! Createspace Independent Publishing Platform, 2018.

[2] D. G. Kleinbaum, Logistic Regression: A Self-Learning Text. Springer Science & Business Media, 2013.

[3] B. Hong, P. Lu, H. Xu, J. Lu, K. Lin, and F. Yang, “Health insurance fraud detection based on multi-channel heterogeneous graph structure learning,” Heliyon, vol. 10, no. 9, p. e30045, May 2024.

4] R. Prabhu, Financial Regulations. Puffins Publishers Private Limited, 2020.

[5] L. G. McDonald, A Colossal Failure of Common Sense: The Inside Story of the Collapse of Lehman Brothers. Crown Pub, 2009.

[6] K. Patukale, Mediclaim and Health Insurance. Prabhat Prakashan, 2021.