NUS
National University
of Singapore

BT4012 Fraud Analytics

AY 22/23 Semester 1

Project Report

Healthcare Provider Fraud Detection Analysis

Group 11	
Clara Tay Linn Qi	A0204413X
Kok Ze Xuan	A0189596B
Loo Hui Lin	A0203151B
Wee Zhen Qi, Tarcius	A0190085E
Wong Chung How, Brian	A0202022J

Table Of Contents

1 Introduction

Fraud is defined as the wrongful or criminal deception intended to result in financial or personal gain. This project aims to identify healthcare providers who may engage in fraudulent insurance claims, by evaluating multiple machine learning models and finding the most appropriate model for this use case.

2 Problem Statement

Healthcare fraud involves healthcare claims being unethically filed to profit illegally from the payouts from insurance companies. It can be committed by healthcare providers, peers of providers and healthcare beneficiaries, with these parties often collaborating to make fraudulent claims. Our project focuses on identifying potential healthcare provider frauds, given the claims they have filed. Common examples of healthcare provider fraud include billing for services not administered, duplicate claims and modifying medical records (Medicare Rights Centre, 2022).

Swift identification of these parties filing fraudulent claims is necessary for insurance companies, as they directly suffer from high estimated losses of nearly $60 billion per year (Thomas, 2022). For Medicare, an insurance program based in the United States that our dataset concerns, 15% of its expenses are due to such dishonest claims. In 2021, it was estimated that there was US$50 billion worth of improper payments from Medicare alone (Federal Government, 2021). Consequently, patients pay higher insurance premiums as insurers pass on the costs of fraudulent payouts to consumers in the form of higher prices. They may also be made to undergo unnecessary medical tests or treatments, putting their health at risk. At the larger scale, healthcare provider fraud burdens the healthcare system, as said providers may compromise on the quality of healthcare that their patients receive. This leads to inefficiencies in the allocation of valuable healthcare resources to patients who require them the most.

Digitally transforming the fraud detection process has been challenging to implement due to the complex and heterogeneous health data systems across the US (Kumaraswamy, 2022). Additionally, the sheer volume of healthcare claims makes fraud detection difficult. For example, Medicare processes more than 4.5 million claims a day (Jimenez, n.d.). The implementation of machine learning models will therefore be useful in alleviating Medicare of the manpower burdens to manually inspect each claim, by providing leads of potential frauds.

3 Data

3.1 Overview of Datasets

This dataset was downloaded from Kaggle and can be accessible from the link: https://www.kaggle.com/datasets/rohitrox/healthcare-provider-fraud-detection-analysis.

It concerns Medicare, an insurance program based in the United States. It mainly serves people over the age of 65 (DCD, 2015).

There were 8 datasets provided in total, 4 each for the train and test set. The test set did not include the target variable `PotentialFraud` so it is not possible to evaluate the models implemented with the test set. Therefore, the test set will be omitted and a new test set will be created using the train dataset provided.

There will be 4 datasets that will be used in this project, namely Train Provider, Train Beneficiary, Train Outpatient and Train Inpatient. Information about the datasets is detailed below.

Train Provider [5410, 2]

This dataset contains provider ID which can be used to merge with both inpatient and outpatient datasets. It also contains the target variable that we are interested in `PotentialFraud`, identifying whether the claim is potentially a fraudulent one.

Data Type	Variables
Object (2)	`Provider`, `PotentialFraud`

Train Beneficiary [138556, 25]

This dataset contains mainly demographical data of the beneficiary. In addition, information on the reimbursement and deductible amounts for both inpatient and outpatient are also included.

Data Type	Variables
Object (4)	`BeneID`, `DOB`, `DOD`, `RenalDiseaseIndicator`,
Integer (21)	`Gender`, `Race`, `State`, `County`, `NoOfMonths_PartACov`, `NoOfMonths_PartBCov`, `ChronicCond_Alzheimer`, `ChronicCond_Heartfailure`, `ChronicCond_KidneyDisease`, `ChronicCond_Cancer`, `ChronicCond_ObstrPulmonary`, `ChronicCond_Depression`, `ChronicCond_IschemicHeart`, `ChronicCond_Osteoporasis`, `ChronicCond_rheumatoidarthritis`, `ChronicCond_stroke`, `IPAnnualReimbursementAmt`, IPAnnualDeductibleAmt`, `OPReimbursementAmt`, `OPDeductibleAmt`

Train Outpatient [517737, 28]

This dataset contains information on the claims of patients who were not admitted to the hospital. It also includes information on the provider, physicians involved, reimbursed amount as well as the amount paid by the patient, which is the claim amount minus the reimbursed amount.

Data Type	Variables
Object (20)	`BeneID`, `ClaimID`, `ClaimStartDt`, `ClaimEndDt`, `Provider`, `AttendingPhysician`, `OperatingPhysician`, `OtherPhysician`, `ClmAdmitDiagnosisCode`, `ClmDiagnosisCode_1`, `ClmDiagnosisCode_2`, `ClmDiagnosisCode_3`, `ClmDiagnosisCode_4`, `ClmDiagnosisCode_5`, `ClmDiagnosisCode_6`, `ClmDiagnosisCode_7`, `ClmDiagnosisCode_8`, `ClmDiagnosisCode_9`, `ClmDiagnosisCode_10`, `source`
Integer (1)	`InscClaimAmtReimbursed`
Float (7)	`DeductibleAmtPaid`, `ClmProcedureCode_1`, `ClmProcedureCode_2`, `ClmProcedureCode_3`, `ClmProcedureCode_4`, `ClmProcedureCode_5`, `ClmProcedureCode_6`

Train Inpatient [49474, 31]

This dataset contains information on the claims of patients that were admitted to the hospital. The information recorded in this dataset is similar to that of the Outpatient dataset. However, it is notable that it also includes key information

Data Type	Variables
Object (23)	\`BeneID\`, \`ClaimID\`, \`ClaimStartDt\`, \`ClaimEndDt\`, \`Provider\`, \`AttendingPhysician\`, \`OperatingPhysician\`, \`OtherPhysician\`, \`AdmissionDt\`, \`ClmAdmitDiagnosisCode\`, \`DischargeDt\`, \`DiagnosisGroupCode\`, \`ClmDiagnosisCode_1\`, \`ClmDiagnosisCode_2\`, \`ClmDiagnosisCode_3\`, \`ClmDiagnosisCode_4\`, \`ClmDiagnosisCode_5\`, \`ClmDiagnosisCode_6\`, \`ClmDiagnosisCode_7\`, \`ClmDiagnosisCode_8\`, \`ClmDiagnosisCode_9\`, \`ClmDiagnosisCode_10\`, \`source\`
Integer (1)	\`InscClaimAmtReimbursed\`
Float (7)	\`DeductibleAmtPaid\`, \`ClmProcedureCode_1\`, \`ClmProcedureCode_2\`, \`ClmProcedureCode_3\`, \`ClmProcedureCode_4\`, \`ClmProcedureCode_5\`, \`ClmProcedureCode_6\`

3.2 Database Schema

The target variable, PotentialFraud, is only available in the Provider dataset. The Provider dataset is joined to the Beneficiary dataset with a 1:N relationship. Inpatient & Outpatient Data linked to Beneficiary Data through Beneficiary ID & Beneficiary Data is linked to Provider Data through Provider ID

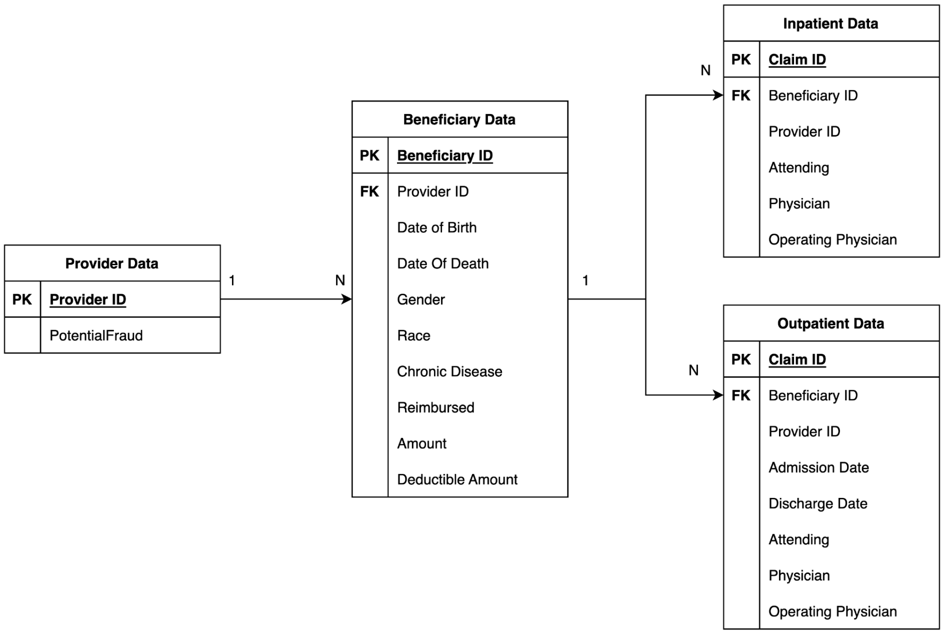

Figure 1: Database Schema

4 Data Preprocessing

4.1 Test Set Creation

Because only the Train data provided comes with the target variables, the test set must be split from the Train dataset to evaluate the final model. The datasets were merged according to their common keys to give a dataset of 558211 observations with 55 features including the target variable of `PotentialFraud`. Thereafter, the test set was split using GroupShuffleSplit() from sklearn.model_selection package, grouping by `Provider`. After splitting the test set, the datasets were broken down into their constituents to allow for neater Exploratory Data Analysis.

4.2 Exploratory Data Analysis & Feature Engineering

Train Provider

With 2 variables in the dataset, the distribution of the class labels were plotted and it was found that only roughly 10% of cases were flagged as potential frauds with 90% being negative. This means the dataset is largely imbalanced.

Train Beneficiary

We used the demographic information to explore whether there were differences between the profiles of fraudulent and non-fraudulent claims. However, there were no differences between them in terms of age and gender. For deductible and reimbursement amounts, we suspected that there may be differences across age and gender, however, the distribution was similar. Some States had significantly more number of beneficiaries than others. The top 3, according to their State Code was Arkansas (5), Delaware (10), New Hempshire (33). These States have a larger percentage of population over the age of 65 as compared to the country average (US Census Bureau, 2021), hence this is not surprising. As for the number of months of coverage, more than 99% of all beneficiaries had 12 months coverage for both, which is the norm since it is part of Original Medicare, the basic protection plan. A ‘Patient Health Score’ was created to provide an overall view of a beneficiaries’ health which could be a better link to monetary amounts as health insurance payouts are highly dependent on one’s health. While there were a few exceptions where lower health scores received higher amounts, most supported this. There was also little to no correlation between the different chronic conditions, other than Renal Disease and Kidney Disease, which both relate to damage of the kidneys.

Train Inpatient and Outpatient

The Inpatient dataset contains information on patients who were admitted into the hospital while the Outpatient dataset contains information on patients who visited the hospital but were not admitted. The Inpatient dataset contains 3 additional columns: Admission Date, Discharge Date and Diagnosis Group Code. We plotted the distribution for the different physicians, CLM diagnosis and procedure codes as well as the diagnosis group code to see if there was any difference between fraudulent and non-fraudulent claims. For physicians, we discovered that certain physician IDs only appeared in fraudulent claims which could be a good predictor. However given that it is user specific data, it does not help in predicting future cases. Instead, it can be used to build a dictionary of fraudulent physicians which the hospital can closely monitor. For CLM diagnosis codes, CLM procedure codes and Diagnosis Group Codes, we did not notice any difference between the distribution of fraudulent and non-fraudulent claims. We tried to identify Physician - Beneficiary pairs which had a high number of occurrences due to potential collusion to perform fraud, however there were no clear differences identified. Finally, we created a variable "Admission Duration" to capture the length of a hospital stay.

5 Modelling

There were several constraints to consider when building our models. Firstly, the cost of false negatives is very high as it leads to unnecessary financial loss and could be detrimental to the health institution and its shareholders. Therefore, we maximise recall in our model optimization function. Next, choosing an interpretable model is important because justifying the prediction of a fraud will assist investigators during their manual inspections. Consequently, black box models like neural networks will be less suitable for this problem compared to interpretable models like decision trees and logistic regression.

We used Bayesian optimization for hyperparameter tuning as it learns from past results, reducing the number of trials required before reaching the optimal set of hyperparameters. This is especially useful given that we have a large number of variables leading to a longer training process. For all models except the Neural Network, the Bayesian optimization was coupled with a stratified two-fold cross validation process to search for the best set of hyperparameters that gave the best recall score. This is done to reduce the overfitting of the model on the training data, thus increasing its generalizability. We decided to do cross validation over 2 folds as our training data is large and therefore should be easily generalizable without the need for too many folds, which would increase the training time. As the output for the neural network model is a probability, we decided to set the threshold where the precision is equivalent to predicting all test cases as fraud, 0.4. We then set values larger than 0.4 to be potential fraud and values lower than 0.4 to be non-fraudulent claims. From this set of predicted fraudulent transactions, we then calculate the respective scores for comparison.

5.1 Logistic Regression

Logistic Regression is a supervised classifier algorithm that determines the likelihood an event will succeed or fail and is used when the target variable is binary (i.e. Fraud / No Fraud). It analyses the relationship from the given dataset and learns a linear relationship before using the Sigmoid function to introduce non-linearity to predict the probabilities.

Advantages

Logistic Regression is generally easy to implement and interpret. It can also be used as a means to find the weight of each feature in predicting the outcome. One can identify the more significant features that help to predict a fraudulent claim by looking at the weight values.

Disadvantages

Issues arise when data is non-linear or has high dimensionality (GeeksforGeeks, 2022). Logistic Regression assumes linearity between the dependent variable and the independent variables, however, this assumption may not hold true for all parameters that we have in our dataset. The relationships may be complex, which is not suitable for Logistic Regression.

5.2 Decision Tree

We also employed a Decision Tree for the binary classification problem, which is constructed by recursively splitting the training samples using the features from the data that work best for the task.

Advantages

It is a white box model which is simple and has high interpretability. Insurers and their fraud investigators can hence easily understand the model decisions, trust the model better, hence apply their final decisions more confidently. Decision tree is also a non-parametric algorithm, hence it does not require the dataset to be a specific distribution (Kapil, 2022). This benefits our training as we have parameters that are largely skewed and distributed differently.

Disadvantages

Decision Trees are unstable (CFI Team, 2022). As a result, a slight change in the data can cause a major change in the decision tree structure. Since the claims are filed by a person, it is prone to human error and thus may contain incorrect data, which can eventually affect the decision outcome.

5.3 Random Forest Regressor

To improve the performance of the single Decision Tree model, we built a Random Forest model, which is a tree-based algorithm that leverages the power of multiple decision trees for making decisions (Sharma, 2022).

Advantages

Using a Random Forest over a Decision Tree may be better suited for the large number of parameters that we have in our dataset.

Disadvantages

Unlike Decision Trees, Random Forests are more like a black box that is hard to understand. Its computations may continue to be more complex, decreasing its interpretability, which as mentioned previously, is an important criteria for choosing the right prediction model. Users should be able to identify how a claim was determined fraudulent.

5.4 XGBoost

XGBoost is a gradient boosted tree algorithm engineered to be more efficient than other implementations of gradient boosting.

Advantages

XGBoost algorithm utilises the gradient boosting framework and is advantageous over other gradient boosting techniques due to its usage of parallel processing which is appropriate for the high dimensionality of the dataset.

Disadvantages

In our business context, determining whether a claim is fraudulent needs to be a fair process. Therefore, it is important that we can interpret the results of the prediction to know how the claim was determined fraudulent. This is one drawback of XGBoost as the model interpretation may not be intuitive.

5.5 Neural Networks

A neural network is a type of machine learning algorithm that uses interconnected nodes in a layered structure to mimic the human brain (AWS Amazon, 2022). The algorithm learns from mistakes and improves continuously which makes it suitable to solve large, complex problems.

Advantages

The large learning capacity of neural networks to learn complex and non-linear relationships (Mahanta, 2017) benefit our large dataset with many diverse parameters. From our exploration, we were unable to identify any significant patterns to focus on for our modelling process, hence by using a neural network, the model learns the features to look for on its own (Sagina, 2018).

Disadvantages

A Neural Network is prone to overfitting (Goyal, 2021) because they learn millions of parameters from the training stages, which is especially amplified with our large number of features. In our baseline model, we simplified the model by limiting the layers. This can be further alleviated by using dropouts or regularisation.

5.6 Model Performance

To evaluate our models, we first compared their recall scores (used to prioritise reducing the rate of false negatives) as this value was maximised during the optimisation processes. However, the cost of false positives should not be neglected as wrongly identifying a genuine claim as a fraud may ruin the insurer's customer relations and cause them to lose credibility. Hence, F1 scores, which considers both the cost of false negatives and false positives, are also compared due to their robustness in evaluating the model performances. They are also helpful in cases of imbalanced datasets.

Listed in the table below are the recall, F1 scores and hyperparameters of the finalised models based on the prediction results on the test data, rounded to 3 significant figures.

Model	**Recall Score**	**F1 Score**	**Hyperparameters**	
Logistic Regression	0.401	0.433	fit_intercept	False
			max_iter	4500
			penalty	l2
Decision Tree	0.272	0.362	criterion	gini

			max_depth	140
			max_features	None
			min_impurity_decrease	0.00206
			min_samples_leaf	10
			min_samples_split	34
Random Forest Regressor	0.135	0.216	criterion	entropy
			max_depth	94
			max_features	None
			min_impurity_decrease	0.00238
			min_samples_leaf	80
			min_samples_split	300
			n_estimators	40
XGBoost	0.314	0.406	colsample_bytree	0.682
			gamma	6.81
			max_depth	18
			min_child_weight	1.0
			n_estimaters	180
			reg_alpha	170.0
			reg_lambda	0.970
			seed	0
Neural Network	0.584	0.464	Number of Layers	3
			Number of Neurons Layer 1	10
			Activation Layer 1	relu
			Number of Neurons Layer 2	20
			Activation Layer 2	relu
			Number of Neurons Layer 3	1

			Activation Layer 3	sigmoid
			learning_rate	0.001

From the table above, we can see that the neural network model performed the best with a recall score of 0.584 and an F1 score of 0.464.

5.7 Integration of Model into Business Process

When a healthcare provider submits a claim, the insurance company can first run the claim information through the model to predict any potential fraud. With this information, domain experts can then selectively evaluate the cases further, hence reducing the total number of cases they have to inspect. This increases the possibility of referrals to the appropriate authorities or recoupments where deserved. Simultaneously, customer relations are strengthened as claim durations reduce due to the improved manpower efficiency.

7 Conclusion

7.1 Limitations

Non-semantic features

As the ClmDiagnosisCode and ClmProcedureCode features are represented by numerical codes in the dataset, we are unable to extract meaningful semantic context from them. The relationship between a beneficiary's diagnosis and the procedure they underwent can inform us of a potentially fraudulent claim, such as when a diagnosis does not include any kidney issues yet the patient underwent dialysis. While this could have been achieved through feature crossings, it would have led to extremely high dimensionality of the dataset (10^{10} additional features) and was therefore unfeasible. Understanding the feature meanings could have enabled a more targeted approach where we simply check for the coherence between the diagnosis and procedure of each claim, therefore creating only one new feature. However, we lacked this information and therefore missed out on a potentially highly predictive feature to identify the frauds.

Computational processing power

Due to limitations in our computational processing power, we had to compromise the number of categories in each variable for encoding by selecting only the top 10 categories. Thus, we lose potential features and information. Furthermore, the hyperparameter tuning for the models were also limited for the same reason, such as reducing the number of epochs and focusing only on selected hyperparameters.

Despite the limitations faced, these are also just present limitations that can be overcomed with potential expansions of the project where more resources are available.

7.2 Future Extensions

Custom ensemble methods

If computational power allows, we can leverage the usefulness of ensemble models. The approach of a custom ensemble method can be summarised as follows: (1) train a selection of different base models, (2) use their predictions to get a new dataset with targets and (3) train a meta-model on this new dataset. This can help to improve the performance and robustness of the fraud detection model.

Application of analysis in related problems

While this project only focuses on healthcare provider fraud, its insights can be extended to other related frauds like beneficiary (receiver of healthcare service) frauds. This is because providers and beneficiaries often work together in the submission of fraudulent claims, such as when providers pay bribes or incentives to knowing patients in return for their cooperation in the fraudulent activity (accepting unnecessary referrals or drugs for non-FDA-approved uses). Understanding the connectivity patterns between these parties in a graph-based approach can therefore extend the model's predictions to beneficiary frauds based on the nature of their interactions. For example, to also flag a patient as suspicious on the premise of their numerous interactions with a fraudulent provider.

References

(DCD), D. C. D. (2015). *What is the difference between Medicare and Medicaid?* HHS.gov. Retrieved from https://www.hhs.gov/answers/medicare-and-medicaid/what-is-the-difference-between-medicare-medicaid/index.html#:~:text=Medicare%20is%20an%20insurance%20program,for%20hospital%20and%20other%20costs

Advantages and disadvantages of logistic regression. GeeksforGeeks. (2022, August 23). Retrieved from https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/

CFI Team. (2022, November 4). *Decision tree*. Corporate Finance Institute. Retrieved from https://corporatefinanceinstitute.com/resources/data-science/decision-tree/

Federal Government. (2021). *Annual improper payments datasets*. Payment Accuracy Page. Retrieved from https://www.paymentaccuracy.gov/payment-accuracy-the-numbers/

Goyal, C. (2021, June 12). *Guide to prevent overfitting in Neural Networks*. Analytics Vidhya. Retrieved from https://www.analyticsvidhya.com/blog/2021/06/complete-guide-to-prevent-overfitting-in-neural-networks-part-1/

Jimenez, A. (n.d.). *Healthcare fraud during a pandemic: Fast facts for financial institutions by Alison Jimenez*. Healthcare Fraud During a Pandemic: Fast Facts for Financial Institutions. Retrieved from https://legal.thomsonreuters.com/en/insights/articles/healthcare-fraud-during-a-pandemic

Kapil, A. R. (2022, October 1). *Advantages and disadvantages of Decision Tree in machine learning*. Blogs & Updates on Data Science, Business Analytics, AI Machine Learning. Retrieved from https://www.analytixlabs.co.in/blog/decision-tree-algorithm/#Interpretability

Kumaraswamy, N., Markey, M., Ekin, T., Barner, J., Rascati, K. (2022, January). *Healthcare Fraud Data Mining Methods: A Look Back and Look Ahead*. Perspectives in health information management. 19. 1i.

Mahanta, J. (2017, July 10). *Introduction to neural networks, advantages and applications*. Towards Data Science. Retrieved from https://towardsdatascience.com/introduction-to-neural-networks-advantages-and-applications-96851bd1a207

Medicare Rights Center. (2022, October 31). *Medicare fraud*. Medicare Interactive. Retrieved from https://www.medicareinteractive.org/get-answers/medicare-fraud-and-abuse/medicare-fraud-and-abuse-overview/fraud-defined#:~:text=Medicare%20fraud%20occurs%20when%20someone,other%20providers%2C%20and%20Medicare%20beneficiaries

Sagina, I. J. (2018, April 25). *Why go large with data for deep learning? | by Ida Jessie Sagina ...* Towards Data Science. Retrieved from https://towardsdatascience.com/why-go-large-with-data-for-deep-learning-12eee16f708

Sharma, A. (2022, June 15). *Decision Tree vs. Random Forest - which algorithm should you use?* Analytics Vidhya. Retrieved from https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/

Thomas, D. L. (2022, April 13). *What is healthcare fraud?* What is Healthcare Fraud? Retrieved from https://www.news-medical.net/health/What-is-Health-Care-Fraud.aspx

U.S. Bureau of Labor Statistics. (2005, September 27). *Appendix D - USPS state abbreviations and FIPS codes*. U.S. Bureau of Labor Statistics. Retrieved from https://www.bls.gov/respondents/mwr/electronic-data-interchange/appendix-d-usps-state-abbreviations-and-fips-codes.htm

The University. (2022). *What is a Neural Network?* AWS Amazon. Retrieved from https://aws.amazon.com/what-is/neural-network/

US Census Bureau. (2021, October 8). *National Demographic Analysis Tables: 2010 (revised)*. Census.gov. Retrieved from https://www.census.gov/data/tables/2012/demo/popest/revised-2010-demographic-analysis-estimates.html

Your coverage options. Medicare. (n.d.). Retrieved from https://www.medicare.gov/basics/get-started-with-medicare/get-more-coverage/your-coverage-options