Health care insurance fraud detection

# Business problem and results

Fraudulent health care insurance claims have significant costs to both health insurance providers as well as consumers. In order to mitigate these costs, we sought to create a model that would help us accurately identify fraudulent claims as well as identify patterns in these claims that would allow us to detect and prevent future fraud. We hypothesized that we could use a K-Nearest Neighbor Model for fraud detection by identifying the circumstances under which providers would be more likely to file fraudulent claims.

Using this model, we found that three factors were most important in determining whether or not there was fraud occurring: the number of days spent in the hospital, the insurance claim amount reimbursed, and the number of attending physicians.

# Background

The National Health Care Anti-Fraud Association (NHCAA) estimates that health care insurance fraud costs the United States approximately $68 billion annually — about 3 percent of the nation's total health care expenditures.[[1]](#footnote-1) Health care fraud leads to higher premiums and higher out-of-pocket expenses for consumers, as well as reduced benefits or coverage. For employers, health care fraud increases the cost of providing insurance benefits to employees and increases the overall cost associated with doing business. The patients themselves may be subjected to unnecessary or unsafe medical procedures, or have their medical records compromised and used to submit falsified claims.

Some common types of health care fraud include:[[2]](#footnote-2)

* Billing for services that were never performed
* Billing for more expensive services or procedures than were actually provided - "upcoding"
* Performing medically unnecessary services for the purpose of accumulating insurance payments
* Misrepresenting non-covered treatments as medically necessary covered treatments for the purposes of accumulating insurance payments
* Billing for each step of a procedure as if they are separate procedures - “unbundling”

In a notable case from 2019, Walgreens agreed to pay a total of $296.2 million in settlements for two separate healthcare fraud cases. The settlement was issued to resolve allegations that it improperly billed Medicare, Medicaid, and other federal healthcare programs for hundreds of thousands of insulin pens that it knowingly dispensed to beneficiaries who did not need them.[[3]](#footnote-3)

Because of the subjective nature of health care and complexity of the American health care system, health care insurance fraud has traditionally been difficult to diagnose.

# Information about the dataset

This dataset was obtained from Kaggle and contains hundreds of thousands of records. At a high level, the data can be broken down into several different categories. First, there is a training dataset that contains records that include potential fraud cases, as well as a test dataset. Within these datasets there are further distinctions:

* Beneficiary data
  + Contains customers details taking Insurance and includes demographic information, geographic information, any pre-existing conditions, and details about insurance policy such as the deductible and amount reimbursed
* Inpatient data
  + Contains information about patients who were admitted to the hospital and has information about the claim, doctors were involved in treatment, insurance providers were involved.
* Outpatient data
  + Contains patients who visited the hospital but were not admitted for treatment and contains information analogous to that contained in the inpatient file
* Provider data
  + Containing the details of whether a provider is making fraud or non-fraud claims.

# Methodology

The datasets contain nearly 50 variables for each claim, of which we need to identify the most effective ones for predicting the fraudulent cases. We tried several different models, including a decision tree, random forest and K-nearest neighbor, as well as different parameters. While running those models, a set of common features can be found useful in making predictions. Below is a detailed description of the variables we selected:

Feature Selection

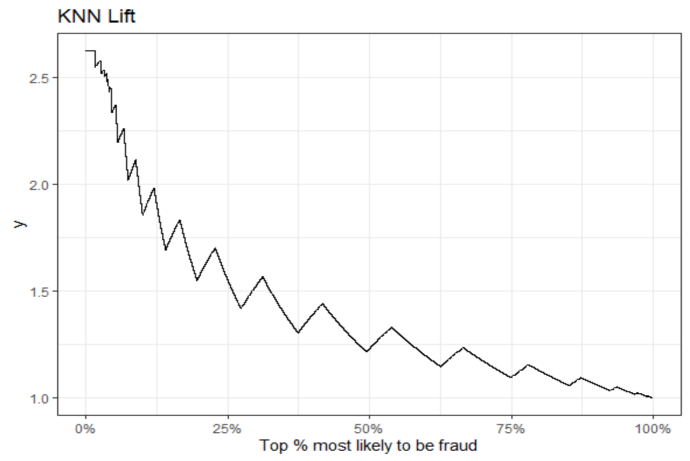
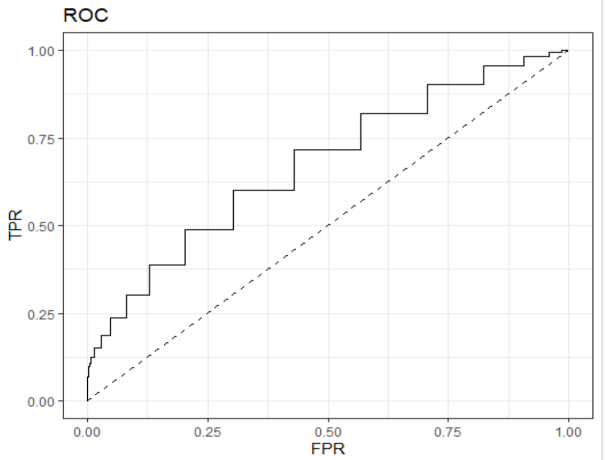
* InscClaimAmtReimbursed: Claim amount reimbursed by the insurance
* DeductibleAmtPaid: Deductible amount paid by the patient
* In\_hospital: Hospitalization days
* Type: whether it is an inpatient or outpatient
* RenalDiseaseIndicator: whether a renal disease is involved
* Disease chronic condition indicator: whether the disease is diagnosed as chronic or not.
* IPAnnualReimbursementAmt: annual amount reimbursed by the insurance as an inpatient
* IPAnnualDeductibleAmt: annual deductible amount paid as an inpatient.
* OPAnnualReimbursementAmt: annual amount reimbursed by the insurance as an outpatient
* OPAnnualDeductibleAmt: annual deductible amount paid as an outpatient.
* Dna: Number of missing diagnosis code involved in a claim
* Physician\_flag: a list of physicians that are highly likely to be involved in fraudulent activities and show up frequently
* Claim\_admit\_diag\_flag: a list of Claim Admission Diagnosis codes that are likely to be related to fraudulent activities

In terms of model building and selecting, we decided to use the F-Score as our performance metric because first, the dataset is imbalanced and second, we want to have a high number of predicted fraudulent cases.

At the same time, we want the model to capture more true fraudulent cases - we want to balance recall and precision, while using the F-score to measure model performance. After trying several different models, we decided to use the K-NN because it gave the best performance (see below). Essentially, the prediction of a claim using the K-NN model is based on the majority fraud status of its nearest 21 neighbors. Given the fact that the dataset is highly imbalanced and non-fraudulent cases greatly outnumber the fraudulent claims, we believe that this model is generally doing a good job capturing potential fraud.

Model performance metrics:

* F-score: 0.40035
* Precision of fraud:  0.5905622
* Recall of fraud: 0.3030797
* AUC: 0.6318



# Key findings

Three factors were most important in determining whether or not there was fraud occurring: the number of days spent in the hospital, the insurance claim amount reimbursed, and the number of attending physicians.

* Fraudulent claims can be costly for both insurance providers and customers. By targeting the top 25% of records that are most likely to be fraudulent, the model performs 1.5 times better in capturing cases of potential fraud than without using a model.
* Without using a model, investigating claims requires a huge amount of effort from the insurance fraud investigation team. But using our model can help identify potential cases of fraud at the first sign before a payment is made and avoid having to wait weeks or months for an adjuster to review a claim, thereby reducing claim handling costs and in some cases avoiding loss costs outright.
* In our sample dataset, with current model and a cutoff level of 0.7, we estimate that a health insurance provider can save about $14,000,000 out of $88,000,000, which is roughly 16% of the total claimed reimbursements.
* By saving a huge amount of money from fraud claims, insurance companies will be able to lower the insurance premium and attract more customers. Thus, health insurance providers will not be able to reduce the cost from fraudulent claims but also increase the revenue by having a larger customer base.

# Limitations to the model:

* Class imbalance: Only a small percentage of claims are actually fraud, so the class distribution of data is highly imbalanced. This leads to decreased predictive performance in classifying the claims which are fraud.
* The model performs a good job on Inpatient data but not on outpatient data. This is because Outpatient data is insufficient to make predictions or draw any patterns as the features which indicate fraud may not have been captured in the existing data. Appropriate data needs to be collected for patients not admitted to hospital and analyzed to find the factors indicating fraud claims from healthcare providers.
* The pattern of fraud can change over time. Healthcare providers can change their patterns of fraud, once they realize that the model has become proficient in detecting their fraud patterns. So, the model needs to be constantly updated to make sure it captures different fraud patterns.

1. https://www.bcbsm.com/health-care-fraud/fraud-statistics.html [↑](#footnote-ref-1)
2. https://www.nhcaa.org/resources/health-care-anti-fraud-resources/the-challenge-of-health-care-fraud.aspx [↑](#footnote-ref-2)
3. https://healthpayerintelligence.com/news/walgreens-agrees-to-296m-settlement-in-healthcare-fraud-cases [↑](#footnote-ref-3)