ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

ΜΕΘΟΔΟΙ ΑΝΑΛΥΤΙΚΗΣ ΤΩΝ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΣΤΑΤΙΣΤΙΚΗΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΣΤΗΝ ΑΝΙΧΝΕΥΣΗ ΤΗΣ ΑΠΑΤΗΣ ΣΤΗΝ ΑΣΦΑΛΙΣΗ ΥΓΕΙΑΣ

Ουρανία Αντωνοπούλου

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Εφαρμοσμένη Στατιστική*

Πειραιάς
Σεπτέμβριος 2022

Π Α Ν Ε Π Ι Σ Τ Η Μ Ι Ο Π Ε Ι Ρ Α Ι Ω Σ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

ΜΕΘΟΔΟΙ ΑΝΑΛΥΤΙΚΗΣ ΤΩΝ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΣΤΑΤΙΣΤΙΚΗΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΣΤΗΝ ΑΝΙΧΝΕΥΣΗ ΤΗΣ ΑΠΑΤΗΣ ΣΤΗΝ ΑΣΦΑΛΙΣΗ ΥΓΕΙΑΣ

Ουρανία Αντωνοπούλου

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Εφαρμοσμένη Στατιστική*

Πειραιάς
Σεπτέμβριος 2022

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. ……. συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της Επιτροπής ήταν:

- Αναπληρωτής Καθηγητής Μπερσίμης Σωτήριος (Επιβλέπων)
- Καθηγητής Γεωργακέλλος Δημήτριος
- Καθηγητής Κούτρας Μάρκος

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIRAEUS

School of Finance and Statistics

Department of Statistics and Insurance Science

POSTGRADUATE PROGRAM IN
APPLIED STATISTICS

DATA ANALYTICS AND MACHINE LEARNING METHODS FOR HEALTH INSURANCE FRAUD DETECTION

By

Ourania Antonopoulou

MSc Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece
September 2022

d

Στη μητέρα μου

Πόπη

Ευχαριστίες

Με την ολοκλήρωση της μεταπτυχιακής διπλωματικής μου εργασίας, θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες στον επιβλέπων καθηγητή μου, κύριο Σωτήριο Μπερσίμη, για την εμπιστοσύνη που μου έδειξε εξ' αρχής, την επιστημονική του καθοδήγηση, τις υποδείξεις του, τη συμπαράστασή του, τη συνεχή του υποστήριξη και το αμείωτο ενδιαφέρον που έδειξε από την αρχή μέχρι το τέλος.

h

Περίληψη

Η έκταση, η πιθανότητα και η πολυπλοκότητα της βιομηχανίας υγειονομικής περίθαλψης έχει προσελκύσει εκτεταμένη απάτη που έχει σημαντικό αντίκτυπο στην οικονομία. Οι δόλιες δραστηριότητες όχι μόνο συμβάλλουν στην αύξηση του προβλήματος των δαπανών της υγειονομικής περίθαλψης, αλλά επηρεάζουν επίσης την υγεία των ασθενών. Η πρόκληση στα τρέχοντα συστήματα εντοπισμού απάτης έγκειται κυρίως στην κατανόηση της επιβάρυνσης της χρηματικής απώλειας και των ασυνήθιστων συμπεριφορών.

Παρά την εφαρμογή διαφόρων τεχνολογιών και στρατηγικών για την καταπολέμηση της απάτης, όπως προγραμματισμένοι, στοχευμένοι και τυχαίοι έλεγχοι, καταγγελίες και βιομετρικά συστήματα, η απάτη στην ασφάλιση συνεχίζει να αποτελεί μεγάλο πρόβλημα για τους περισσότερους παρόχους ασφάλισης υγείας.

Σκοπός της παρούσας διπλωματικής εργασίας είναι να ευαισθητοποιήσει σχετικά με αυτόν τον σημαντικό τομέα το ευρύτερο κοινό αλλά και να καταδείξει την αναγκαιότητα για τους σχετικούς δημόσιους οργανισμούς και τις εταιρείες του χώρου της υγείας να επενδύσουν ένα σημαντικό μέρος του κεφαλαίου τους αλλά και του χρόνου τους στη σοφή αξιοποίηση των δεδομένων, τα οποία διατίθενται πλέον στον κόσμο, στην στατιστική μηχανική μάθηση, καθώς και σε τεχνικές προβλεπτικής αναλυτικής με σκοπό την μείωση της απάτης στην ασφάλιση υγείας.

Αρχικά, δίνονται οι σχετικοί ορισμοί για την απάτη στην ασφάλισης υγείας καθώς και μία επισκόπηση του προβλήματος και των τρόπων και μορφών με τις οποίες αυτή εμφανίζεται. Στη συνέχεια, υπογραμμίζεται με διάφορα παραδείγματα απάτης στον τομέα της υγειονομικής περίθαλψης πόσο σημαντικό είναι το πρόβλημα. Επιπλέον, παρατίθενται και άλλοι τομείς εντοπισμού απάτης, οι οποίοι υποστηρίζονται με διάφορες μελέτες περιπτώσεων. Στη συνέχεια γίνεται μία διεξοδική αναζήτηση στη βιβλιογραφία για τις εφαρμογές της ανάλυσης δεδομένων και της στατιστικής μηχανικής μάθησης που έχουν χρησιμοποιηθεί μαζί με μελέτες περιπτώσεων. Έπειτα, αναφέρονται οι συνήθεις μέθοδοι μηχανικής μάθησης στον εντοπισμό της απάτης. Τέλος, επιλεγμένες μεθοδολογίες εφαρμόστηκαν σε πραγματικά δεδομένα προκειμένου να παρουσιαστεί μια ολοκληρωμένη μελέτη περίπτωσης με στόχο τη δημιουργία

ενός μοντέλου που μπορεί να εντοπίσει με χαμηλότερο κόστος και δαπάνες δόλιους παρόχους στην ασφάλιση υγείας.

Λέξεις – Κλειδιά

Απάτη; απάτη στην ασφάλιση υγείας; μηχανική μάθηση; Αναλυτική δεδομένων; ανίχνευση απάτης; εποπτευόμενες μέθοδοι; μη εποπτευόμενες μέθοδοι; υβριδικές μέθοδοι; περιοχές ανίχνευσης απάτης.

Abstract

The extent, probability, and complexity of the healthcare industry have attracted widespread fraud that has a significant impact on the economy; fraudulent activities not only contribute to the problem of increasing health care costs but also affect patient health; the challenge in current fraud detection systems lies mainly in understanding the burden of financial loss and unusual areas of behavior.

Despite the implementation of various anti-fraud technologies and strategies, such as planned, targeted and random checks, complaints, and biometric systems, insurance fraud is still a major problem for most health insurance providers.

The purpose of this thesis is to raise awareness of this important area among the general public and to demonstrate the need for relevant public bodies and health companies to invest a significant part of their capital and their time in the wise exploitation of the data now available in the world, statistical machine learning, as well as predictive analysis techniques to reduce health insurance fraud.

Initially, the relevant definitions for health insurance fraud are given, and it is provided an overview of the problem and the ways and forms in which health insurance fraud occurs. Next, it is highlighted how important the problem is with various examples of healthcare fraud worldwide, and a presentation of punishments for health insurance fraud. In addition, other fraud detection areas are listed, supported by case studies. Then there will be a thorough search in the literature for the applications of data analytics and statistical machine learning that have been used along with case studies. At a later stage, the methods of machine learning in fraud detection that are commonly used are mentioned. Finally, selected methodologies were applied to real-world data in order to present a complete case study that aims to build a model that can detect fraudulent providers with lower costs and expenditures.

Keywords

Fraud; health insurance fraud, machine learning; data analytics; fraud detection; supervised methods; unsupervised methods; hybrid methods; areas of fraud detection.

Table of Contents

r

List of Tables

List of Figures

List of Acronyms

ACA	Affordable Care Act
ANN	Artificial Neural Networks
ARP	Accounts Receivable Pipelines
AUC	Area Under ROC curve
CLT	Central Limit Theorem
CMS	Centers for Medicare and Medicaid Services
CNN	Convolutional neural network
EFD	Electronic Fraud Detection
EHFCN	European Healthcare Fraud and Corruption Network
EPBS	Emergency Physician Billing Services
EU	European Union
FBI	Federal Bureau of Investigation
FCPA	Foreign Corrupt Practices Act
FDiBC	Fraud Detection in Bank Club
FFD	Financial Fraud Detection
FN	False Negative
FNR	False Negative Rate
FP	False Positive
FPR	False Positive Rate
GA	Genetic Algorithm
GDP	Gross Domestic Product
HCC	Healthcare Continuum
HHC	Department of Health and Human Services
HIP	Health Information Pipelines
HIPAA	Health Insurance Portability and Accountability Act
IRDA	Insurance Regulatory and Development Authority
KNN	K-Nearest Neighbor
MLP	Multi-layer perceptrons
MSE	Mean Squared Error
NHCAA	National Health Care Anti-Fraud Association
NHIN	National Health Information Network
NN	Neural Network

NTL	Non-Technical Loss
ONC	Office of the National Coordinator for Health Information Technology
OR	Odds Ratio
PCA	Principal Component Analysis
PHI	Protected Health Information
ROC	Receiver Operating Characteristic
SDEM	Sequentially Discounting Expectation and Maximizing
SEC	Security and Exchange Commission
SMOTE	Synthetic Minority Over-sampling Technique
SOM	Self-Organizing Map
STAT	State Transition Analysis Tool
SVC	Support Vector Classification technique
SVM	Support Vector Machines
TP	True Positive
TPR	True Posivite Rate
TN	True Negative
TNR	True Negative Rate

CHAPTER 1

Introduction

1.1 Insurance

Insurance is a contract (policy) under which a person or a corporation receives financial protection or reimbursement from an insurance company against losses. The company pools the risks of its clients to make payments more reasonable to the insured. Nowadays, customers choose an insurance policy for three reasons: to be protected against risk factors, as an investment, and for tax purposes. Accidents and health difficulties are common today, causing accidental movement in people's lives. An accident will make a massive loss for a human being. Because of this, everyone intends to protect their own life; to do so; they require a safety net. The insurance company helps people who have an accident and suffer health problems.

Nevertheless, at the same time, nowadays, the insurance sector faces more problems due to internal and external side, for example, customer claim frauds in the insurance sector and lack of knowledge of internal auditors. The data is entered into the database manually by the internal auditor, so due to carelessness and lack of knowledge of the internal auditor, some misrepresentation will occur, for example, the date of birth entered after the death date. These things will make an unexpected loss for the insurance sector and the customers. Next, every customer should surrender all the legal documents when joining any insurance sector and provide the proper documents to collect the amount from the insurance sector, which is not happening in real life.

Due to the above reasons, some loss occurs in the insurance sector. Insurance is generally classified into four types mentioned below, and this thesis describes the details of the last, which is health insurance fraud detection.

- Home insurance
- Life insurance
- Motor insurance
- Health insurance.

1.2 Insurance Fraud

Insurance fraud is a willful misrepresentation or falsification that the individual or business makes, knowing that the deception may result in unauthorized use by the person, company, or another party. The most common type of fraud is a fabricated statement, misrepresentation, or willful mistake crucial to deciding the benefits to be provided. Fraudulent acts are always illegal, even though the type and scope of the offences differ from one place to the next.

Over a year, millions of health insurance claims are submitted to healthcare companies helping make quality care affordable. A certain percentage of these millions of claims are fraudulent, costing the system, taxpayers, insurers, and the government billions of dollars annually, leading to higher premiums and other expenses. Dishonest service providers are not the only ones who commit health insurance fraud. Instead, we can see how fraud is growing throughout the healthcare system at all levels and tiers. Such frauds can taint a practitioner's or hospital's reputation and affect the entire healthcare system. Health insurance companies can substantially decrease fraud incidents by using proper medical billing services and obtaining the most effective medical coding services.

1.3 Healthcare Fraud

The health insurance industry continues to have much money flowing through it. There are several implications, especially in prevention, detection, and investigation. These three areas necessitate a thorough understanding of every legitimate and illegitimate player in the Healthcare Continuum (HCC), the ability to identify Health Information Pipelines (HIP) and Accounts Receivable Pipelines (ARP), and a grasp of how all players use Protected Health Information (PHI).

Understanding how the HCC operates and how changes in the healthcare sector affect fraud prevention, detection, and investigation is critical. A tip is the most prevalent way fraudulent activity gets started. The information provided led to the beginning of an investigation of the situation's outcome. A thorough analysis permits an in-depth understanding of the healthcare industry's characteristics. Healthcare fraud is frequently hidden within crucial business activities. The HIP and ARP processes determine which functions should be explored. These are detected as soon as it is figured out what is happening. Investigations and detections will reveal weaknesses, which can be employed as preventative measures. Prevention necessitates an awareness of how the healthcare system works and how the cycle repeats.

According to the "Report on the Use of Health Information Technology to Enhance and Expand Health Care Anti-Fraud Activities", published by the Office of the National Coordinator for Health Information Technology (ONC), the healthcare business is in a position that is startlingly similar to that of the financial services industry 15 years ago. The banking industry was transitioning from a paper-based system to a sophisticated electronic one at the time. The banking industry tackled the inefficiencies

of paper systems by investing extensively in information technology infrastructure, according to a well-thought-out vision and plan. Credit card fraud, which is now believed to be less than 7 cents per $100, is often considered a big issue. However, healthcare fraud is 100 times more costly.

The role of technology is also discussed in the ONC's paper mentioned above. Its creators believe that technology can play an essential role in detecting fraud and abuse and helping to prevent it. Although technology cannot eliminate fraud, it can dramatically reduce fraud and abuse, resulting in lower healthcare fraud losses. Fraud loss reduction requires using advanced analytics tools built into the National Health Information Network (NHIN).

The information available via the NHIN must comply with all federal and state laws. The federal government continues its efforts to identify healthcare fraud, waste, and abuse continue to grow. Healthcare firms must implement an effective compliance program. It is particularly essential to develop a corporate culture that fosters ethical behavior. Many healthcare organizations are using corporate compliance initiatives to foster such corporate cultures.

1.4 Health Insurance Fraud

Health insurance systems are funded by governments or administered by the private sector to share healthcare costs in these countries. According to the National Health Care Anti-Fraud Association (NHCAA), healthcare fraud is intentional deception or misrepresentation by an individual or entity that could result in an unauthorized benefit to them or their accomplices.

The health insurance fraud characteristics include damage level insufficient information, a suspected diagnosis of proof, insured low willingness to cooperate and cause of the accident unreasonable, repeatedly claims record, in a particular area, occur at a specified time for late submitting claims. Inconsistent application documents, high claim payments, certificates of poor reliability, non–cooperation are very familiar with insurance knowledge. An affiliate files each claim with the approval of a physician justifying incapacity. Data received by the affiliate in the last three months and included in each form consist of age, gender, type of claim, name, and date of birth of affiliate, ID number, requested rest period, type and location of rest, healthcare professional identification, employer identification, work activity of the company where the affiliate works, affiliate occupation and income records. As health insurance is national-wide, it takes 45 days minimum to complete the process. So, the research is to solve workforce and make the proper decision-making using data analytics technology.

The most common type of health insurance fraud is false claims. The aim is to receive an unearned payment for a series of claims. So, to make health insurance viable, one must focus on eliminating or reducing fraudulent claims.

There are two types of insurance fraud. The first is complex fraud, which is a deliberate attempt either to point out an event or an accident that requires hospitalization or another type of loss that would be covered under a medical insurance policy. The second is soft fraud, which can also occur when people

purposely provide false information regarding the pre-existing illness or other relevant information to influence the subscription process in favour of the applicant. Such as claim fraud, application fraud, and eligibility fraud.

1.5 Impact of Healthcare Fraud

1.5.2 International Cases of Healthcare Fraud

1.5.2.1 Fraud in the United States

Examples of healthcare fraud are plentiful. In an insurance company, all payments for foreign claims are made to the insureds and not to foreign medical providers. An insured patient submitted fictitious foreign claims, as per $90,000, from a clinic in South America, indicating that the entire family was in a car accident. A fictitious police report accompanied the medical claims. A telephone call to the clinic revealed that the insured and their dependents were never treated there.

Because she worked at a medical facility, the woman had access to claim forms and patients' medical records. She claimed that she had more than 27 surgeries, operations on her gall bladder, finger amputations, and a hysterectomy, among other procedures. The intention was to cash in on the service checks as quickly as possible. The oddity was that whenever a patient had surgery, a bill for the matching hospital stay should have been turned in; nevertheless, this was not the case.

An employer who colludes with applicants to receive benefits illegally or commits fraud to avoid taxes will be penalized with a minimum of $500 and should even be prosecuted. For instance, "collude" means knowingly helping applicants obtain benefits to which they are not entitled, for instance, cash wages or other hidden compensation for services performed. In other words, the employer misrepresents the applicant's eligibility so that he or she will receive benefits if he or she is not qualified.

Emergency Physician Billing Services, Inc. (EPBS), a third-party medical billing company, provides coding, billing, and collections services for emergency physician groups in over 100 emergency departments throughout 33 states. Based on allegations presented by a qui tam relator, a whistleblower reporting fraud, the United States charged that EPBS and its principal owner, D J. D. McKean, routinely billed federal and state healthcare programs for higher levels of treatment than were provided or supported by medical history documentation. EPBS was compensated for its assistance in the engagement of billed or recovered revenues, depending on the customer. In a second case, a supply vendor-provided adult diapers, which Medicare does not cover, mistakenly billed them as expensive prosthetic devices called "female external urine collecting devices." In the third instance of vendor fraud, an ambulance company charged ambulance rides for shopping outings.

A third-party administrator processing claims on behalf of Medicare signed a company integrity agreement with the Department of Justice in response to a variety of allegations by providers that the third-party administrator did not process claims consistent with coverage determinations, did not process

or pay physicians' or other healthcare claims in a timely fashion, or at all. Also, he applied incorrect payments for appropriate claims submissions and inaccurately reported claims processing data to the state. Failed to provide home health services to qualified beneficiaries, automatically changed current procedural terminology codes to explain the procedure provided, did not recognize modifiers, and did not reliably answer patient appeals, sometimes not responding at all or waiting 6 to 12 months.

1.5.2.2 Cases of Fraud in Other Countries

Healthcare fraud is widespread not only in the United States but also in other countries around the world. All below examples include patterns of behavior consistent.

- South Africa: In South Africa, in 2004, a newspaper reported that "a man who posed as a homoeopathic doctor was this week sentenced to 38 years in jail, the harshest sentence ever handed out by a South African court for stealing from medical aids."
- Australia: By falsifying patient referrals and charging for time spent in intimate interactions with patients, an Australian psychiatrist could claim more than $1 million.
- United Kingdom: A medical researcher in the United Kingdom misled his colleagues and the public by using his urine sample for 12 research subjects. Switzerland, famous for its watches, has providers authorized to bill for 30-hour days.
- Greece (The Smith and Nephew scandal, 1998 - 2008): Smith and Nephew is a global medical device firm that sells orthopaedic, endoscopic, and wound-care equipment. It has facilities all over the world. The US Security and Exchange Commission (SEC) has charged the London-based medical device company, Smith and Nephew, with violating the Foreign Corrupt Practices Act (FCPA) for bribing public doctors in Greece to win business for more than a decade. Smith and Nephew subsidiaries devised a strategy to pay bribes to Greek doctors through a web of offshore firms and subsidiaries, including US and German subsidiaries, beginning in 1997. Charges alleged that Smith and Nephew channelled from 1998 to 2008 more than 9 million US dollars to persuade Greek surgeons to use its artificial hips and knees. The Greek distributor of Smith and Nephew justified the bribery system, saying that competitors were paying even higher rates at the time. In February 2012, the US subsidiary of Smith and Nephew agreed to pay more than 22 million US dollars (about 17 million euros) to the SEC and Ministry of Justice. Smith and Nephew's chief executive commented: 'These legacy issues do not reflect Smith and Nephew today. Nevertheless, they underscore that we must remain vigilant in every place we do business and let nothing compromise our commitment to integrity.
- Austria (Research subsidy, 2006): In March 2006, a case at the Innsbruck University Clinic of Traumatology and Sports Medicine made headlines. A supplier of prosthetic parts made a payment of 57,000 euros into the account of the Association for Research into Trauma Surgery in Tirol. The article accused the management of the University Clinic of accepting the money in exchange for research into product improvements made for the company. The money was returned, and the State Attorney dropped

the case. The management announced that in the future, all externally sourced research funding would be paid into a central research fund and distributed following a transparent procedure and all the rules.

- Hungary (Bribery for delivery, 2012): In January 2012, eleven gynaecologists and two midwives were accused of repeatedly asking patients to pay for delivery in MAV Hospital's (State Health Centre) maternity ward. The hospital brought a charge against their doctors (involving the former Head of Department). According to the accusation/indictment, they asked for money in at least 20 cases for the procedure (for conducting the birth, anaesthesia, and analgesia) between April 2007 and July 2008. These procedures are provided free of charge in the social health insurance package. The prosecutors asked for a financial penalty for the doctors and one midwife and a prison penalty for the other midwife accused of bribery committed on a commercial scale. The accused denied that the mothers had requested the money. So far, the media has not reported any follow-up on the case.

1.5.3 The Financial Cost of Healthcare Fraud

1.5.3.1 The Financial Cost of Healthcare Fraud in the United States

Insurance fraud has a significant impact on the healthcare system. Healthcare fraud ultimately leads to higher premiums, consumer spending, and a loss in benefits or coverage, regardless of whether the person has employer-sponsored health insurance or self-insured coverage. Healthcare fraud, which affects private and governmental enterprises, raises the cost of providing insurance benefits to employees and the total cost of conducting business. The enormous quantities involved in fraud may determine whether or not health insurance becomes a reality for many individuals. On the other hand, financial losses due to healthcare fraud are only half of the picture.

Unfortunately, victims of healthcare fraud are easy to come by. These people have been duped, manipulated, and forced to undergo needless or dangerous medical operations. Patient records are hacked, and valid insurance information is occasionally used to create fraudulent claims. Patients with private health insurance frequently have a lifetime limit or other benefit restrictions under their coverage. The monetary amount counts for the rest of the patient's life or limits each time an incorrect claim is paid on their behalf. This means that the limit may have been reached even if a patient has a legitimate need for insurance coverage. Healthcare has long been regarded as a recession-proof business, and it continues to expand. The following statistics are staggering in their implications.

- According to the Centers for Medicare and Medicaid Services (CMS), originally known as the Healthcare Financing Administration, Americans spent about $42 billion on healthcare in 1965. In 1991, the money spent had risen to $738 billion, a 1,657 per cent increase. By 1994, US healthcare consumers spent $1 trillion. The money spent in 2004 had risen to $1.6 trillion, or $6,280 per healthcare customer. By 2008, the value exceeded $2.2 trillion, or nearly $250 million each hour. A trillion-dollar market contains around $329.2 billion in fat or roughly 25 per cent of yearly spending. Also, based on the

Centres for Medicare and Medicaid Services, $108 billion, or 16 per cent, of the above is paid improperly due to billing errors.

- Furthermore, CMS counts that the prescription drug market is $121.8 billion annually, making the annual counterfeit price tag approximately $12.2 billion.
- According to National Centre for Policy Analysis, $33 billion of Medicare dollars, or 7 per cent, are illegitimate claims billed to the government.
- According to MBA news, $100 billion in private-pay dollars, or 20 per cent, are estimated to be paid improperly.
- $50 billion, or 10 per cent, of private-payer claims, are paid out fraudulently, based on Blue Cross/Blue Shield.
- $37.6 billion is spent annually on medical errors, based on Agency for Healthcare Research and Quality.
- Based on Food and Drug Administration, ten per cent of drugs sold worldwide are counterfeit, up to 50 per cent in some countries.
- The United States now spends about $2.6 trillion annually on health care, which is 17.5 per cent of GDP (Gross Domestic Product). The proposed reform attempts under the Affordable Care Act (ACA) will drastically increase the number of Americans covered and the amount spent, leading to even more fraud, waste, and abuse in the system. Healthcare fraud is a widespread issue that affects federal, state, and private insurance programs. Healthcare fraud has increased dramatically in the United States during the last decade, with billions of dollars paid out on false claims.
- Health care fraud accounts for 3 per cent of all healthcare spending, or $60 billion, according to the National Health Care Anti-Fraud Association. Other estimates put the figure at around $200 billion.
- According to the FBI, fraudulent billings to healthcare systems account for between 3 per cent to 10 per cent of overall healthcare spending, totalling annual financial losses between $19 billion and $65 billion.
- Medicare and Medicaid made an estimated $68.3 billion incorrect payments in 2010. The property or casualty industry also participates in the successful Healthcare Fraud Prevention Partnership. In this way, a collaborative effort occurred in which Medicare, Medicaid, TRICARE, the VA, private health plans, and others shared information and data on shady medical providers.
- Medicare accounts for 20 per cent of all US healthcare spending with a total possible cost recovery, with the potential application of effective fraud detection methods, of $3.8 to $13 billion from Medicare alone. To date, this public-private cooperation has saved more than $300 million, according to the US Government Publishing Office.
- The World Health Organization’s latest estimate of global healthcare expenditure in the US is $5.7 trillion, or €4.13 trillion. Thus, it is likely that around the US $415 billion, or €301 billion, is lost

globally to fraud. This money was enough to build more than 2,300 new hospitals, measured at developed world prices.

- According to the Coalition Against Insurance Fraud, insurance fraud causes $80 billion in damage to American consumers yearly. The cost of healthcare fraud is estimated to be $54 billion annually in the United States. When an insurance fraud occurs, it affects not just the insurance firm but also ordinary consumers.
- In 31 federal districts across the country, the Department of Justice announced criminal charges against 138 people, including 42 doctors, nurses, and other medical professionals. The allegations stem from the defendant's alleged involvement in various healthcare fraud schemes that resulted in nearly $1.4 billion in losses. Telemedicine fraud costs $1.1 billion, COVID-19 healthcare fraud costs $29 million, fraud involving drug addiction treatment centres costs $133 million, and illegal opioid distribution and other healthcare fraud costs $160 million in the United States.

These statistics mean that about $25 million per hour is stolen in healthcare in the United States alone and show unnecessary healthcare expenses that directly influence the cost and quality of healthcare for all Americans. Healthcare fraud and abuse not only raise insurance rates, but each dollar spent on false or abusive claims diminishes the amount of money available to enhance the quality of treatment for individuals with valid medical bills.

1.5.3.2 The Financial Cost of Healthcare Fraud in Other Countries

In recent years, the number of empirical research on insurance fraud coupled with studies on market failures, information asymmetry, and poor regulatory measures in financial sectors of economies across the globe is increasing continuously. This has come because of massive losses attributed to insurance fraud on the global insurance markets, which run into billions of dollars, affecting the growth of insurance firms and the financial well-being of both insured and uninsured. Numerous studies have shown that healthcare costs in the United States are growing faster than the overall economy and globally, as the following statistics show.

- Insurance Europe states in its European Insurance Anti-Fraud Guide from 1996 that the cost of fraud in the European insurance market cannot be less than €8 billion, or roughly 2% of total yearly premium income from all classes of insurance combined.
- In 2008 health care expenditure consumed slightly under 9 per cent of GDP in the European Union (EU), which is over 1 per cent above the health sector's share of GDP a decade before. While much of this spending has gone toward necessary health care goods and services, a significant proportion, potentially up to 30 per cent, may have been lost to wasteful spending.
- Another example of European abuse is the European Health Insurance Card (EHIC), which only permits urgent and unplanned healthcare for those travelling abroad, but patients travel to obtain healthcare. From 2000-to 2004, according to European Healthcare Fraud and Corruption Network

(EHFCN), €2.5 billion related to cross-border care was left unpaid between the Member States of the European Economic Area, and many of these transfers involve fraud but are left undetected and unresolved.

- Furthermore, the EHFCN estimates that €56 billion is wasted annually in the EU due to healthcare fraud, equating to nearly €80 million per day. In 2008, Denmark's loss estimate was the highest at €13,016 million, while Malta's was the lowest at €23 million.
- In 2009, an EHFCN survey found 4,188 suspected fraud incidents in six countries, with the Netherlands and Belgium reporting the most significant numbers (1,075 and 2,884 cases, respectively). A total of 3,875 (93 %) of the cases discovered were investigated further, with 469, or 11 per cent, being referred for prosecution. Approximately one-quarter of those recommended for the prosecution have been successfully prosecuted.
- A Belgian dentist stole €1 million and was sentenced to prison as an example of provider fraud. He improperly billed nearly 200 patients whose contact information he took from a database he had access to while working with two other dentists between 2000 and 2008.
- A supplier of healthcare goods is a Spanish company that was found to deliver inferior quality wheelchairs to patients at a discounted price, even though the wheelchairs did not match the brand that the doctors ordered. The company director was convicted of fraud and sentenced to imprisonment, and his contact with civil services was cancelled. In addition, he had to pay €23,775 in damages to the Catalan Health Inspectorate.
- The extent of fraud detected in the healthcare fraud sector was €46.3 million in France, €43 million in Germany, €18.7 million in the Netherlands, €11.9 million in the United Kingdom, €6.8 million in Belgium, €0.8 million in Lithuania, and €0.3 million in Greece, based on EHFCN for the fiscal year 2016.
- The Canadian Coalition Against Insurance Fraud (CCAIF) estimates from a study conducted in 1997 that $1.3billion worth of general insurance claims paid in Canada every year is fraudulent. In 2013, Canadian Coalition Against Insurance Fraud estimated that this amount has increased by 5 to 10 per cent.
- In 1997, in Australia, 10 per cent of all insurance premiums paid by the public were lost to fraud, with the total amount paid out for fraudulent claims each year running to $1.4billion, based on Baldock.
- In 2015, in South America and the Caribbean, insurance fraud is estimated to cost between 19 and 35 per cent of the annual revenue of the insurance industry based on Fraud Intelligence.
- In 2015, in South Africa, 100 million rands were lost in 2010 because of policyholder/claim or consumer fraud, based on South Africa Insurance Crime Bureau.

- In 2014, in Kenya, Kuria and Moronge posited that 40 per cent of the insurance claims paid by insurers are fraudulent, and, in Nigeria, in 2011, it was estimated that between 10 and 30 per cent of insurance claims submitted are fraudulent, based on Yusuf.

1.6 Punishments of Health Insurance Fraud

Fraud in the healthcare system is punishable civilly and criminally under federal law. Penalties such as fines, jail, and a payback order are included in the distinction between civil and criminal sanctions (compensation of the victim for any money lost through fraud). Civil penalties are imposed only for restitution and not for imprisonment or fine. Crime fraud in the health sector can have severe consequences for all convicts at the federal and federal levels, which are mentioned below.

Prison: Healthcare fraud is a severe crime that can result in lengthy prison sentences. 5-year imprisonment for each offence of making a false or misleading statement on a Medicaid or Medicare claim, whereas a federal conviction for federal health fraud can result in a 10-year sentence for each violation. A healthcare-related fraud that results in serious personal injury to a person can result in a term of up to 20 years in prison, while a healthcare-related fraud that results in death can result in a life sentence.

Fines: Individuals convicted of healthcare fraud face heavy fines. For example, a person who makes a false statement in a Medicaid or Medicare claim might face a punishment of up to $250,000, while organizations that make fraudulent claims could face a fine of up to $500,000. Organizations that take part in ongoing programs that handle healthcare breaches in various ways could face fines of millions, if not billions, of dollars.

Restitution: As part of the criminal sanctions, the judge may order the defendants to refund the money they received illegally because of their fraudulent activity. A doctor who has not been adequately billed to an insurance company and paid for these tests may be required to return that money to the insurance company. The reimbursement comes on top of a fine paid to the government.

Probation: If you are guilty of a health crime, you may be eligible for a suspended sentence. Instead of putting someone in prison, the probationary term restricts their liberties. Probationary periods usually are at least 12 months long, although they can go up to three years. Probationers must adhere to specific requirements, such as meeting with a probation officer regularly, maintaining employment, having no ties to known criminals, and not committing any more crimes. Health insurance fraud is a severe crime with consequences for all parties involved, including government officials and taxpayers, insurers and premium payers, healthcare providers, and patients. This is something that all stakeholders must keep in mind if the system is to be successful.

1.7 Chapter Summary

The first chapter introduces the principal concepts of fraud in the healthcare sector. Drawing an analogy with the problem of fraud in the healthcare sector, it introduces many central definitions, such as insurance, insurance fraud, healthcare fraud and health insurance fraud. This chapter also gives an overview of the impact of healthcare fraud, including cases of fraud and financial cost of fraud worldwide, as also punishments for both civilly and criminally under federal law.

CHAPTER 2

Application Domains and Classification of Fraud

2.1 Types of Health Insurance Fraud

Health insurance fraud is classified as follows by the Insurance Regulatory and Development Authority (IRDA):

Claim fraud: Fraud committed against insurance companies is during the purchase and execution of an insurance product, including deception committed while filing a claim. In other words, the client is claiming health coverage to which he/she may not be entitled per the policy's terms and conditions.

Application Fraud: In this case, the customer inputs incorrect information despite knowing all of the facts about his pre-existing terms, conditions, date of birth, claims, and so on. Put another way, policyholders with pre-existing diseases (such as diabetes) do not reveal them to save money on premiums and gain more coverage. There are those situations where we cannot help ourselves.

Eligibility Fraud refers to supplying incorrect information to gain access to benefits and, for example, disclosing the employee's incorrect job status to achieve eligibility for a health insurance policy with more benefits. To put it another way, let us assume that a part-time employee fabricates employment documents to be qualified for the company's health insurance.

Medical Identity Theft: In this case, a policyholder's name is incorrectly utilized to gain access to free medical services or treatments without the policyholder's agreement or knowledge. False insurance claims are filed in these situations to obtain free care. As a result, fraudulent data is entered into the policyholder's medical records.

External and internal frauds: External fraud refers to deception perpetrated against the insurance business by the client, service provider, or beneficiaries. Internal fraud is committed against the customer by agents or insurance companies.

It is critical to separate health frauds from simple mistakes, omissions, or improper payments. To commit fraud, a person knowingly must exercise a plan, program, or activity to provide falsehoods to make a financial gain. A mistake that results in a patient being billed for therapy he did not receive is not the same as fraud. In contrast, a healthcare professional who deliberately offers treatments or

procedures the provider knows the patient does not require and then bills an insurer for these procedures to generate a profit is considered a healthcare fraud.

Healthcare fraud differs from healthcare abuse. Abuse refers to:

- Incidents or procedures that do not adhere to the standard of care (substandard care)
- Unnecessary costs to a program, caused either directly or indirectly
- Improper payment or payment for services that fail to meet professional standards
- Medically unnecessary services
- Substandard quality of care (for example, in nursing homes)
- Failure to meet coverage requirements

Healthcare fraud, on the other hand, can take many forms, and it can be performed by either a doctor or a consumer. Some forms of healthcare fraud perpetrated by providers include:

- Billing for services that were never performed.
- Falsifying a patient's diagnosis to justify the use of non-medically necessary tests, surgeries, or other procedures.
- Misrepresenting procedures to get paid for non-covered services like cosmetic surgery.
- Upcoding refers to billing for a more expensive service than the one provided.
- Unbundling refers to billing each procedure stage as a separate procedure to maximize reimbursement.
- Accepting bribes (in kind or cash) in exchange for patient referrals.
- Excess charging the insurance carrier or benefit plan by waiving patient copays or deductibles.
- Billing a patient for services that were prepaid or reimbursed in full by the patient's coverage plan for more than the copay amount.

Consumer health insurance fraud can take many forms, including:

- Obtaining medical services or prescriptions using a forged or expired identity card.
- Giving a medical identification card to someone not eligible to use it.
- Adding a non-eligible individual(s) to a contract for coverage.
- Falsifying or altering medical bills and receipts.

2.2 Other types of fraud

2.2.1 Financial Fraud

Eighteen firms were used as a study reference by academicians from U.S.A and China to study and analyze data analytics techniques' performance in fraud detection. Chang and C.-J. Lin recommended a general data analytics financial fraud detection framework, as depicted in Figure 1.

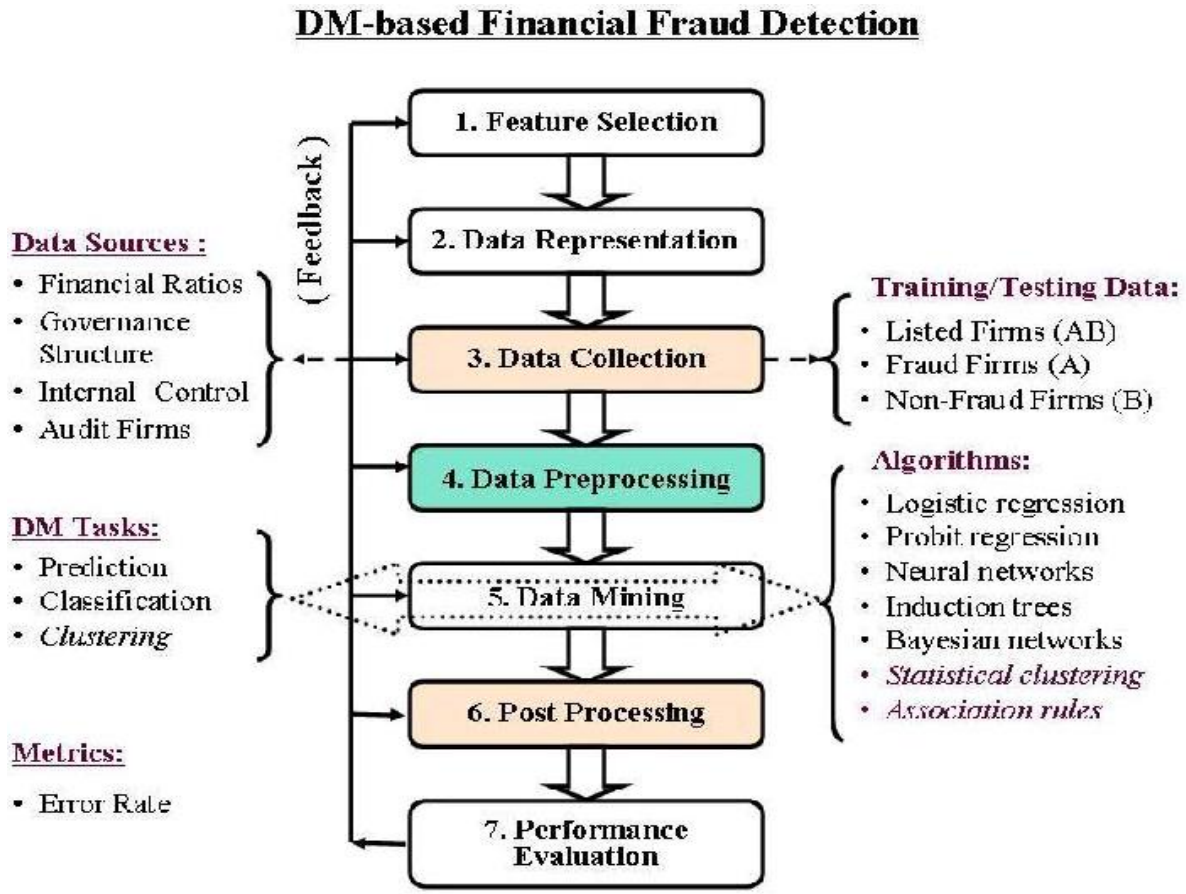

Figure 1: General Data Analytics Financial Fraud Detection Framework

2.2.1.1 Bank Fraud

According to Connell University Law School, bank fraud is defined as “whoever knowingly executes, or attempts to execute, a scheme or artifice to defraud a financial institution; or to obtain money, funds, credits, assets, securities, or other property owned by, or under the custody or control of, a financial institution under false or fraudulent pretenses, representations, or promises.”

Credit card fraud, money laundering, and mortgage fraud all fall under the umbrella of bank fraud, where mortgage fraud is defined as "An underwriter or lender relies on a material misstatement, misrepresentation, or omission relative to the property or potential mortgage to fund, purchase, or insure a loan." and credit card fraud is defined as the transactions on an inactive card, the use of the unauthorized card, or unusual transaction behavior. According to the Federal Bureau of Investigation (FBI), money laundering is the process by which criminals conceal or disguise the proceeds of their crimes or convert them into goods and services. It allows criminals to infuse their illegal money into the economy, corrupting financial institutions and the money supply while giving criminals unjustified economic power. Gao and Ye similarly define money laundering as the process by which criminals "wash dirty money" to disguise its illicit origin and make it appear legitimate and "clean."

2.2.1.2 Insurance Fraud

Insurance fraud can occur at many points in the insurance process (for instance, application, eligibility, rating, billing, and claims) and can be perpetrated by consumers, agents and brokers, insurance company personnel, healthcare professionals, and others. As already mentioned, insurance fraud includes agricultural, healthcare, and automotive insurance fraud. FBI argues that healthcare fraud is carried out by several sectors of the healthcare system through multiple methods, with some of the most frequent types of fraud including “Billing for Services not Rendered, Upcoding of Services, Upcoding of Items, Unbundling, Duplicate Claims, Excessive Services, Medically Unnecessary Services, and Kickbacks". Crop insurance fraud is performed by purchasers of crop insurance who fabricate or misrepresent either the loss of their crops due to natural disasters or the loss of revenue due to reductions in the price of agricultural commodities. Automobile insurance fraud encompasses fraudulent practices, including manufactured accidents, needless repairs, and falsified personal injuries.

2.2.1.3 Securities and Commodities Fraud

The FBI provides brief descriptions of some of the most general securities and commodities frauds encountered today, for example, "Market Manipulation, High Yield Investment Fraud, The Ponzi Scheme, The Pyramid Scheme, Prime Bank Scheme, Advance Fee Fraud, Hedge Fund Fraud, Commodities Fraud, Foreign Exchange Fraud, Broker Embezzlement, and Late-Day Trading." According to another definition by Cornell University Law School, securities fraud consists of theft from manipulating the market, theft from securities accounts, and wire fraud.

2.2.1.4 Other Related Financial Fraud

The final category is constituted of types of financial fraud other than those in the categories mentioned earlier, such as corporate fraud and mass marketing fraud. Again, based to the FBI, “corporate fraud investigations include the following three activities, falsification of financial information, self-dealing by corporate insiders, and obstruction of justice designed to conceal any of the above-noted types of criminal conduct.”. “Mass marketing fraud is a general term for types of fraud that exploit mass-communication media, such as telemarketing, mass mailings, and the Internet”, as the Bureau states.

2.2.1.5 Techniques for Identifying Financial Fraud

Some available techniques for identifying financial fraud are discussed as follows.

Classification: Classification sets up and applies a model to anticipate the categorical labels of unknown objects to distinguish between objects of various classes. These categorical labels are established, discrete and unordered. Zhang and Zhou state that classification and prediction establish a set of standard features and models that describe and distinguish data classes or ideas. Common classification approaches include neural networks, the naïve Bayes, decision trees, and support vector machines. Such classification tasks are employed in the detection of credit cards, healthcare and

automobile insurance, and corporate fraud, among other types of fraud, and classification is one of the most prevalent learning models in data analytics in FFD (Financial Fraud Detection).

Clustering: Clustering is used to divide objects into conceptually significant groups (clusters), with the objects in a group being similar but substantially dissimilar to the objects in other groups. Clustering is also known as data segmentation or partitioning and is viewed as a variation of unsupervised classification. According to Yue et al., "clustering analysis covers the challenge of decomposing or partitioning a data set (usually multivariate) into groups so that the points in one group are comparable to each other and are as dissimilar as possible from the points in other groups." Further, Zhang and Zhou claim that each cluster is a collection of data objects within the same cluster but unlike those in other clusters. The most general clustering approaches are the K-nearest neighbour, the Naïve Bayes methodology, and self-organizing map techniques.

Prediction: Prediction estimates numeric and ordered future values based on the patterns of a data set. Han and Kamber remark that, for prediction, the characteristic for which the values are being predicted is continuous-valued (ordered) rather than categorical (discrete-valued and unordered) (discrete-valued and unordered). This characteristic can be referred to simply as the expected attribute. Neural networks and logistic model prediction are the most often used prediction techniques.

Outlier detection: Outlier detection is performed to measure the “distance” between data objects to detect those objects that are drastically different from or inconsistent with the remaining data set: “Data that appear to have different features than the rest of the population are called outliers”. Yamanishi et al. point out that the topic of outlier/anomaly detection is one of the most fundamental challenges in data analytics. A typically utilized strategy in outlier detection is the discounting learning algorithm.

Regression: Regression is a statistical method used to uncover the relationship between one or more independent variables and a dependent variable (continuous-valued). Many research studies have utilized logistic regression as a standard. The regression methodology is often conducted using such mathematical methods as logistic regression and linear regression, and it is employed in detecting credit cards, crop and automobile insurance, and corporate fraud.

Visualization: Visualization refers to the easily understood presentation of data and to a methodology that transforms complex data characteristics into clear patterns to allow people to observe the complex patterns or relationships revealed in the data analytics process. Eick and Fyock explain that researchers at Bell and AT and T Laboratories have used the pattern detection capabilities of the human visual system by designing a suite of tools and applications that flexibly encode data using colour, position, size, and other visual features. Visualization is best used to provide complex patterns through clearly presenting data or functions.

2.2.2 Computer Intrusion

2.2.2.1 Categorizing Computer Intrusion Fraud

Intrusion is the possibility of an intentional unauthorized attempt to access information, manipulate information, or render a system unreliable or unusable. Intruders may be an outsider (or hacker) and an insider who knows the system's layout, where the valuable data is, and what security safeguards are in place. Two distinct categories can be applied to computer intrusions: misuse intrusions and anomaly intrusions. Misuse intrusions are well-defined attacks on areas of a system that are already known to be vulnerable. Identifying anomaly incursions requires making observations of usage patterns that deviate from the system's norm. These include unauthorized entry attempts, attacks using masquerades, leaks, denial of service, and malicious use of the system.

Many intrusion detection systems base their operations on analyzing audit data generated by the operating system. An audit trail is a record of things to do on a system that is logged to a file in chronologically sorted order. An intrusion detection system is needed to automate and execute system monitoring by keeping aggregate audit trail statistics. Intrusion detection approaches can be classified into two categories based on the model of intrusions, misuse and anomaly detection.

Misuse detection attempts to recognize the attacks of previously observed intrusions in the form of a pattern or a signature (for example, frequent changes of the directory or attempts to read a password file) and directly monitor for the occurrence of these traits. Misuse techniques include expert systems, keystroke dynamics monitoring, state transition analysis, and model-based reasoning. Since unique attack sequences are incorporated into the abuse detection system, known attacks may be detected reliably with a low false alarm rate. Misuse detection is simpler than anomaly detection. However, a significant disadvantage of misuse detection is that it cannot anticipate all the different attacks because it only detects recognized patterns of abuse.

Anomaly detection intends to create an average historical profile for each user and then utilize significant divergences from the profile to flag suspected intrusions. Anomaly detection approaches include statistical methods, predictive pattern generation, and neural networks. The advantage of anomaly detection is that it can detect novel attacks on systems because it evaluates current activity against statistical models for historical behavior, not bed with specific or pre-defined patterns. However, there are some inabilities of this approach. It is expected to have high rates of false alarms. Unusual but legitimate use may occasionally be considered abnormal. Statistical measures of the user profile can be gradually trained, so intruders can train such systems over a period until intrusive behavior is considered normal. Also, it cannot identify the specific type of attack that is occurring. Moreover, anomaly detection systems are computationally expensive because of the expense of keeping track of and updating various system profile metrics.

2.2.2.2 Techniques for Identifying Computer Intrusion

The techniques utilized in misuse detection and anomaly detection are listed as follows.

Expert Systems: An expert system is a computing system capable of describing and reasoning about some knowledge-rich domain to solve issues and give advice. Expert system detectors encode information about assaults as if-then rules. NIDES, developed by SRI, uses the expert system technique to implement an intrusion detection system that performs real-time monitoring of user activities. NIDES consists of a statistical analysis component for anomaly detection and a rule-based analysis component for misuse detection.

Neural Networks: Neural Network Intrusion Detector is an anomaly intrusion detection system constructed by a backpropagation neural network under a UNIX environment. It is trained to recognize users based on commands and how often they use them daily. It is easy to train and affordable because it operates offline on daily log data. ANN (Artificial Neural Networks) provides the capacity to generalize from previously observed behavior (normal or malicious) to anticipate similar future unseen behavior for anomaly identification and abuse detection. It is implemented with a hack propagation neural network.

Model-based Reasoning: Model-based detection is a misuse detection technique that detects assaults by observable behaviors that derive an attack signature. There is a database of attack scenarios with a sequence of behaviors making up the attack. Garvey and Lunt linked theories of misuse with evidential reasoning. The system builds up evidence for an intrusion attempt until a certain threshold is reached, at which point it flags an intrusion attempt. Kumar and Spafford propose a pattern-matching strategy based on Colored Petri Nets to detect abuse infiltration. Under the UNIX environment, audit trails are used as input.

Data Analytics: Data analytics methods can be employed for intrusion detection. A vital advantage of the data analytics approach is that it can construct a new class of models to detect new threats before human specialists have detected them. A classification model with an association rules algorithm and frequent episodes is created for anomaly intrusion detection. This approach may automatically construct succinct and accurate detection models from a significant quantity of audit data. However, it takes much audit data to compute the profile rule sets.

Moreover, this learning process is an integral and continuous part of an intrusion detection system because the rule sets used by the detection module may not be hestatic over a long period. A team of researchers at Columbia University presented the detection models utilizing cost-sensitive machine learning methods. Audit data is examined using an association rules algorithm to determine static features of attack data.

State Transition Analysis: State Transition Analysis is an abuse detection technique in which attacks are described as a sequence of state transitions of the monitored system. Actions that lead to intrusion scenarios are defined as transitions between states. Intrusion scenarios are represented in terms of state

transition diagrams. Nodes reflect system states, while arcs represent appropriate actions. If a compromised (final) condition is ever reached, an incursion is considered to have occurred. STAT (State Transition Analysis Tool) is a rule-based expert system developed to seek known penetrations in the audit trails of multi-user computer systems.

2.2.3 Telecommunication Fraud

2.2.3.1 Categorizing of Telecommunication Fraud

Fraud is costly to a network carrier in terms of lost profit and wasted capacity. There are two types of telecommunications fraud: subscription fraud and superimposed fraud.

Subscription fraud happens when a person obtains a subscription to a service, frequently using a false identity, with no intention of paying. This topic also includes cases of bad debt. Superimposed fraud happens when a service is used without the necessary authorization, as evidenced by the appearance of unknown calls on a bill. Mobile phone cloning, ghosting (the technology that tricks the network to obtain free calls), insider fraud, tumbling (rolling incorrect serial numbers are used on cloned handsets so that repeated calls are credited to various legal phones), and other methods are examples of this fraud.

2.2.3.2 Techniques for Identifying Telecommunication Fraud

Previous work in detecting telecommunication fraud has primarily focused on detecting overlaid fraud. Most strategies use Call Detail Record data to develop consumer behavior profiles and detect deviations from these profiles. These approaches are discussed as follows.

In the rule-based method, a combination of absolute and differential usage is validated against specific rules and matched to data in toll tickets. Flexible criteria for detecting any usage change in a thorough user behavior history can be constructed using differential analysis. The rule-based method works well with user profiles that provide explicit information, and fraud criteria can be called rules. The rule-discovery methodology combines two data levels, which are customer data and behavior data (usage characteristics in a short time frame). A greedy algorithm with altered thresholds is used to pick a ruleset. Siemens ZFE created PDAT, a rule-based intrusion detection tool. PDAT is utilized for mobile fraud detection due to its flexibility and broad applicability.

Rule-based analysis can be complicated to manage because properly configuring such rules requires precise, laborious, and time-consuming programming for each imaginable fraud possibility. The emergence of new fraud types requires constantly updating these rules to include existing, emerging, and future fraud choices. Furthermore, it is a significant impediment to scalability. The more data the system must process, the more severe the performance degradation.

Neural Networks: Neural networks are commonly employed in the detection of fraud. Neural Networks can calculate user profiles independently, allowing them to adapt to different users' behavior more gracefully. It is claimed that neural networks can significantly reduce operating costs. The European Commission's ASPeCT project evaluated the viability of implementations using a rule-based

approach and neural networks, supervised and unsupervised learning, using data from toll tickets. This project presented three approaches based on toll tickets (call records stored for billing purposes). A supervised learning-based feed-forward neural network is utilized for training a nonlinear discriminative function to classify subscribers using summary statistics. Second, density estimation with the Gaussian mixture model is applied to model each subscriber's past behavior and detect any abnormalities in the past behavior. Third, Bayesian networks create probabilistic models based on subscribers' behavior.

Visualization Methods: X-Visualization techniques rely on human pattern recognition to detect anomalies and provide close-to-real-time data feeds. The idea is that while machine-based detection methods are primarily static, the human visual system is dynamic and can quickly adapt to the ever-changing techniques used by fraudsters. Visual data analytics, combining human detection with machines for increased computational capacity, creates a user interface to manipulate the graphical representation of the number of calls made between subscribers in different geographical locations to detect international calling fraud.

2.3 Case Studies

Artificial intelligence distributed and parallel computing, econometrics, expert systems, fuzzy logic, genetic algorithms, machine learning, neural networks, pattern recognition, and visualization are among the critical approaches utilized for fraud detection identified by Bolton and Hand (2002). The practice methods are not often described in detail, and data sets are not generally shared with the public for privacy reasons or to prevent criminals from obtaining vital information.

One systematic way is to compare the transactions to a baseline of expected occurrences and mark the unusual observations for further investigation. These types of outlier identification tools also aid in the detection of new fraud schemes. Such benchmark-based criteria, for example, are commonly used in telecommunications networks to detect anomalous activity.

Another critical set of methodologies involves calculating risk (suspicion) scores for each transaction based on classification or regression model outputs. These are more successful in domains where auditors know the fraud methods. The banking industry, in particular, has had great success in combating fraud using supervised approaches. Ngai et al. (2011)'s review focuses on the use of data analytics tools for financial fraud detection. The abundance of labelled data is one of the primary advantages of financial fraud approaches. Credit card databases contain information about each transaction, including merchant code, account number, type of purchase, client name, size, and transaction date. Banks have identified fraudulent and valid prior transactions and real-time access to this information. This enables the use of supervised methods such as logistic models, neural networks, Bayesian belief networks, and decision trees for classifying financial transactions. Because fraudulent transactions are relatively rare, one should be aware of the imbalanced class sizes. Next, there are some fraud implementations.

Hasheminejad and Salimi (2018) propose a novel sliding time and scores window-based method, FDiBC (Fraud Detection in Bank Club), to detect fraud in bank clubs. n FDiBC, 14 features are produced based on each score obtained by bank club customer members, and then five sliding time and scores window-based feature vectors are proposed based on each customer member's scores. Positive and negative labels are used for generating training and test datasets from the obtained scores of fraudsters and familiar customers in a bank's customers' club system. After generating the training dataset, learning is performed through two approaches: clustering and binary classification with the OCSVM method for positive data, i.e., fraudster customers, and multi-class classification including SVM, C4.5, KNN, and Naïve Bayes methods. The results reveal that FDiBC can detect fraud with 78 per cent accuracy and thus can be used in practice.

Lelenguiya (2015) proposed a model for detecting Non-Technical Loss (NTL) of commercial electricity utilization utility using data analytics technologies like Naïve Bayes, neural network, K-Nearest Neighbor, and Support Vector Machine. He applied data analytics techniques about customer information invoicing or billing systems for electricity utilization in a selection of accounts at Kenya Power Limited. The effectiveness and correctness of the model were verified and assessed to get one accepted technique adopted by Kenya Power Limited. From the results of the tested model, the most significant outcome for fraud detection hit rate is attained by the support vector machine (SVM) classifier with 86.44 per cent, followed by K-Nearest Neighbor with 84.75 per cent, and the classifier with the least optimal fraud detection rate is the Naïve Bayes at 74.58 per cent.

Mamo (2013) investigated the potential suitableness of the data analytics methodologies in building designs and prototypes that can identify and foretell doubtfulness in a levy or tax claims. In his research, he first applied the clustering algorithm to the dataset and then applied classification techniques to develop the predictive model. The K-Means clustering algorithm is applied to uncover the standard classification of the various levy claims as fraudulent or not fraudulent. The next cluster is then applied in the development of the classification model. J48 decision tree and Naïve Bayes classification algorithms were used in this study to build the prototype that can foresee suspicious fraud levy claims best. The model built on the J48 decision tree algorithm displayed the highest classification accuracy of 99.98 per cent. The model was assessed with 2200 testing datasets and recorded a prediction accuracy of 97.19 per cent.

Wei et al. (2013) proposed a novel algorithm to extract variance patterns effectively and determine unscrupulous from actual behavior, backed up by a working and usable pattern choice and risk scoring that integrates predictions from independent design models. The results from investigations on a large scale of accurate online banking data prove that the system can attain higher accuracy and smaller alert volume than the modern benchmarking fraud unearthing systems integrating domain knowledge and conventional fraud discovery techniques.

Eyad (2012) developed an intelligent model that predicts and selects suspicious water bills for discovering fraudulent undertakings. The model improves the discovery hit rate from 1-10 per cent ad hoc manual discovery to 80 per cent intelligent discovery. Support Vector Classification technique (SVC) was applied to uncover the irregular customers' load profile operations.

Ogwueleka (2011) developed a neural network design for the credit card identification system with the use of an unsupervised technique that was used to the transactions data to create four groups of low, high, risky, and high-risk groups. The self-grouping map neural network approach was applied to crack the challenge of bringing out the best groupings of each record to its related group. The Receiver Operating Characteristic (ROC) curve for credit card fraud identification identified over 95 per cent of fraud incidences without prompting any false panics contrary to other mathematical designs and the two-stage clusters.

2.4 Chapter Summary

In chapter 2, we discussed the types of health insurance fraud and the difference between healthcare fraud from healthcare abuse was highlighted as well; other fraud detection areas are listed along with their techniques for identifying fraud, supported by case studies.

CHAPTER 3

Healthcare Fraud: A Literature Review

3.1 Medical Claims Data

3.1.1 The Complex Nature of Medical Claims Data

Despite widespread fraud detection approaches in all areas, medical fraud assessment has received less attention (Phua et al., 2010). Some of the methods mentioned above can be used to detect fraudulent medical claims. Travaille et al. (2011) discuss the applicability of fraud detection methods in other industries, such as credit card, telecommunications, and computer security, to U.S. Medicaid healthcare programs. Medical insurance claims, for example, can be compared to credit card transactions in terms of recording the information of the provider and beneficiary as well as the transaction details.

The heterogeneity and complexity of medical systems and data make the broad application of fraud detection tools in healthcare hard. The nature of medical data differs from that of other industries. It is well mentioned that data do not generally report in real-time or accurately. Labelling data necessitates an audit, making retrieval costly and time-consuming. Furthermore, data confidentiality is critical, and patient privacy is protected by several federal laws, including the Health Insurance Portability and Accountability Act (HIPAA).

Meanwhile, fraud patterns are dynamic and adapt over time to legislation, medical procedures, and billing changes. Because of this dynamic nature and a lack of labelled data, supervised methods are not as usually used as other industries' fraud detection frameworks (Furlan and Bajec, 2008). On the other hand, outlier detection methods are more challenging due to data heterogeneity, the pressure of timely claim processing, and the need for medical expertise in evaluation. For example, the auto insurance industry has strict controls and can halt payments to providers with different billing characteristics. However, in the healthcare industry, timely payments are critical, and unusual provider behavior may be due to medical necessity.

Rising budget deficits in developed countries have focused public attention on healthcare spending. This fact has increased efforts to reduce health care spending by cutting off unnecessary payments via medical audits and fraud assessments. Pre-payment reviews are carried out in a broader sense, mainly through identification checks that can be used to filter out fraudulent transactions (Suleiman et al., 2014).

Despite this, the majority of medical audits are performed after payment. The aim is to identify and recover improper payments through efficient detection procedures (CMS, 2016b).

Each medical investigation requires the topic area competence of specialists who manually audit claims. The expenditures of the study are proportional to the time spent by the expert and the physical resources. Unnecessary auditing of claims with no overpayment (false positives) results in a loss of faith in the government and a cost of opportunity. The heterogeneity of medical claims data and the presence of multiple fraud patterns, on the other hand, can increase estimation error.

One of the most difficult challenges in subsequent resource allocation decisions is the trade-off between audit costs and extrapolation accuracy. Domain experts should ideally identify fraudulent activities through comprehensive medical audits. However, that is generally impractical because of the complex nature and the enormous size of the medical claims data. Descriptive data analysis tools can assist in detecting clear patterns and revealing potential cases of overpayment. It is employed when overpayment can be easily discovered due to irrefutable evidence, such as providers invoicing for beneficiaries who cannot be served.

These challenges necessitate the systematic application of statistical approaches such as sampling and data analytics to assess medical fraud.

3.1.2 Overview of Medical Claims Data

The properties of medical data differ from program to program and country to country. The literature offers examples from the health insurance programs of many countries. This section seeks to provide a high-level, not exhaustive, description.

Generally, the medical data can be categorized as practitioners', medical claims, and clinical instance data (Liu and Vasarhelyi, 2013). Practitioners' data describe service providers in a certain period and describe provider-related characteristics of service cost, quality, and utilization (Viveros et al., 1996). Yang and Hwang (2006) describe clinical-instance data as a set of activities performed by medical staff during a particular treatment. The vast majority of raw medical data is from insurance claims. A medical claim involves the participation of a patient and a service provider and generally incorporates the qualities of patients, providers, and the claim itself. Attributes of a patient can be gender, age, and medical history, whereas the type and the location of the facility are among the attributes of a provider. The prescription specifics, monetary amount, and paid amount are among the crucial characteristics of the claim.

Fraud data contain numerous distinct properties. First, fraud is rare since valid claims nearly always exceed fraudulent ones. For instance, more than 80 per cent of the studies assessed in Phua et al. (2010) include skewed data with less than 30 per cent fraud. The sparsity of the fraud data can be handled with non-negative matrix factorization, principal component analysis, and singular value decomposition (Zhu

et al. (2011)). Secondly, false claims and litigation have dynamic patterns due to increased competition in the health care sector and regulatory and legal framework updates.

Moreover, the fraudulent instances are no longer homogeneous due to the several forms of fraud happening in the same period (Fawcett (2003)). Furthermore, fraudulent situations are not homogeneous because multiple types of fraud co-occur (Fawcett 2003).

In most industry practices, pre-processing data operations, such as transformation and data cleansing, require, most of the time of entire fraud detection method (See Lin and Haug (2006) and Sokol et al. (2001) for related topics). For instance, submitting the same claim under the hospital or provider's name may require the investigators to define new unique identities to examine the medical data (Musal, 2010).

Another major medical data issue is the number of missing values. Missing data can generate problems such as over/under-sampling, non-representativeness, and potential bias in inference. Despite that, many articles in the literature do not specifically disclose how they manage missing data before their data analysis. One of the generally used ways is to eliminate the claim lines with incomplete information. For instance, Ortega et al. (2006) dismissed 35 per cent of these medical claims in a particular year because of poor quality in terms of missing variables and low contribution. By eliminating instances with missing attribute values, Yang and Hwang (2006) filter out noisy data. Removing the instances can diminish the statistical power of analysis since possibly valuable information in the other fields is lost. In addition, the pattern of missing values may be systematic, and deleting records may create a limited selection. However, there are not any systematic rules to handle missing data. Domain specialists should carefully evaluate the benefits and downsides of managing missing values. Grzymala-Busse and Hu (2000) give a comparison of a few techniques. Such methods include replacing missing values with the mode or the mean of the data set, with random values taken from the underlying distribution, or treating missing values as user-defined constants. Imputation, substituting a missing value using an estimate derived by statistical analysis such as regression, might be presented as another means of coping with missing values (Li et al., 2008). Little and Rubin (2014) provide a complete overview of such imputation strategies.

Choosing and altering the properties (features) is also a critical step in data analysis. Attributes utilized in fraud research can be the numerical, category, or binary type of variables. Different types of variables would need the employment of different statistical approaches. The selected features are rarely mentioned in publications due to agreements with the data sources. The objective of such secrecy agreements is to prevent criminals from getting access to how detection systems work. In practice, features are often picked by medical domain specialists. Domain experts are informed about common fraud occurrences or the categories with the highest financial losses.

The relevant attributes of providers for fraud detection are classified into five primary categories: financial, logistics, medical logic, abuse, and identification (Major and Riedinger, 2002). From the investigation perspective, the money amount and the paid fraction are crucial features of a claim. Manual

feature selection can benefit from statistical testing of each picked feature's relevance and importance (See Dash and Liu, 1997) for an overview of feature selection approaches). For instance, examining the link between attributes is typically disregarded in industrial practices while picking attributes for outlier identification approaches. Ortega et al. (2006) are one of the uncommon research projects that utilize correlation checks to eliminate redundant features and examine each feature's discriminating ability.

It is prohibitively time-consuming and costly to evaluate all or most observations. Therefore, many statistical sampling and overpayment estimating approaches are proposed. On the other hand, data analytics technologies are often used because of the heterogeneous structure of data and the nature of fraud.

3.2 Sampling and Overpayment Estimation in Medical Claims Data

In the last decade, there have been multiple attempts to give overviews of various areas of the emerging statistical medical fraud assessment field.

Li et al. (2008) present a significant detailed review of the application of data analytics approaches through 2007. However, it should be noted that recent data sharing and transparency efforts by governmental health organizations have given researchers more access to medical claims data and resulted in more recent publications. This fact has resulted in several overviews, especially aspects of medical fraud assessment. Capelleveen (2013) presents an outline of the utilization of outlier detection methods in medical fraud assessment. Whereas Joudaki et al. (2014) provide a short review of data analytics methods in healthcare fraud detection, Bauder et al. (2017) give a detailed survey on the state of healthcare upcoding fraud analysis and detection with an emphasis on data analytics. All these surveys specialize in data analytics methods, and none of them examines statistical methods like sampling and overpayment estimation.

Sampling and overpayment estimation methods can help medical auditors retrieve the sample data and extrapolate. Various techniques, such as simple random and stratified sampling, are recommended to draw representative samples from the population of interest as efficiently as possible. In the U.S., using probability sampling methods for medical investigations has been accepted as part of the legal framework since 1986. Yancey (2012) presents a full review of these legal sampling procedures and the parties engaged in U.S. public medical insurance programs. There are governmental software packages that assist auditors with sampling and analysis. For instance, in the U.S., medical auditors can use RAT-STATS (OIG, 2010), which the Office of Inspector General offers, Office of Audit Services. RAT-STATS can perform functions such as determining sample size, generating random numbers to select the sample, and providing inference. Woodard (2015) gives a brief review and a simple application to demonstrate the use of sampling by the United States Medicaid program to detect and reclaim overpayments.

One of three outcomes can result from the payment amount associated with a claim. A claim could be classified as totally legitimate, totally illegitimate, or partially overpaid. A claims data set where each claim is either a legal payment or a wholly illegitimate payment is "all or nothing". In most situations, according to the current U.S. sampling guidelines (CMS, 2001), the lower limit of a one-sided 90 per cent confidence interval for the total overpayments should be used as the recovery amount from the provider under investigation. Using the lower bound allows for a reasonable and fair recovery without requiring tight precision to support the point estimate, the sample mean. In other words, the state is protected with confidence from recovering an amount greater than the actual value of erroneous payments. However, this application of the Central Limit Theorem (CLT) assumes that the overpayment population either follows the normal distribution or that the sample size of overpayments is reasonably large. Mohr (2005) illustrates that normality-based models can perform well for cases with one overpayment pattern.

It is common for medical claims data to exhibit skewness and non-normal behavior, requiring large sample sizes to validate CLT. Edwards et al. (2003) illustrate that methods focused on the CLT may not work well for specific overpayment populations with tiny sample sizes. They suggest the "minimum sum method," a nonparametric inferential method that uses the hypergeometric distribution and computes the respective lower bound estimates. These estimates are efficient in settings where the claims are essential, "all or nothing." The payment population is relatively homogeneous and well separated from zero.

Many extensions are proposed for the minimum sum method. Ignatova and Edwards (2008) present a sequential sampling methodology that tries to make inferences on the proportion of claims involving overpayments. Gilliland and Feng (2010) provide an adaptation to address cases of varying payments. It is improved by Gilliland and Edwards (2010) using randomized lower bounds, in which payment amounts are audited in equal-sized packets. Edwards et al. (2003) discuss a simple extension, so-called q-adaptation. The minimum sum method is based on the re-definition of illegal payments, so the payments are considered illegitimate if "q" per cent of the payment is in error. In trying to adapt to partial overpayments as well as "all or nothing" claims, Ekin et al. (2015) present a zero-one inflated mixing model that extends Mohr (2005). Musal and Ekin (2017) present a Bayesian mixture model that can be more efficient for claims with partial overpayments. In addition, standard stratified expansion and combined ratio estimators of the total are among the proposed estimators.

3.3 Analytics Methods

3.3.1 Supervised Algorithms

Supervised methods are based on using labelled fraudulent and non-fraudulent records to classify claims and generate predictions. Regarding classification algorithms, Li et al. (2008) underline the extensive use of neural networks and decision trees for fraud detection. Liou et al. (2008) compare logistic

regression, decision trees, and neural networks in terms of their ability to identify medical fraud effectively.

Neural networks can manage complex, massive data sets and nonlinear variable connections. However, implementing neural networks often requires statistical expertise, such as tweaking the parameters. In addition, Padmaja et al. (2007) point out that classification may demonstrate poor performance and overfitting with skewed data sets. In order to avoid overfitting, Ortega et al. (2006) employ an early stopping technique in their neural network-based medical fraud detection algorithm. It is based on using one training data set to update the weights and biases and another data set to terminate training when the network begins to overfit the data. They also handle a significant prediction variance due to a limited sample size with multiple features.

In comparison, decision trees have generic rules which are easy to interpret, especially with a small number of categories. Decision trees can also handle sparse data but may result in overfitting and decreasing outcomes' interpretability with the growing data size. For instance, Shin et al. (2012) present a scoring model for the likelihood of misuse and then categorize providers using a decision tree.

Ormerod et al. (2003) advocate a dynamic Bayesian network of fraud indicators, with weights, decided using every feature's fraud prediction power. Bayesian classifiers have comparatively shorter training times and are shown to address the data's sparsity successfully. He et al. (1998) employ the k-nearest neighbour technique to classify practitioners' practice profiles. As an alternative, a support vector machine-based approach is applied by Kumar et al. (2010). Medical detection efforts include merging multiple supervised algorithms. Chan and Lan (2001) use the fuzzy sets principle and a Bayesian classifier to detect fraudulent claims in Taiwan National Health Insurance. Viveros et al. (1996) offer a mix of association rules and a neural segmentation algorithm for fraud detection.

Overall, supervised algorithms are beneficial for detecting previously established patterns of fraud. Since they are based on past classified claims, one should be wary of potential overestimation issues (Liou et al., 2008). The availability of unequal class sizes within the claims data might potentially lead to overfitting. These models should be continually updated to deal with new fraud tendencies and changes in legislation. The incapacity of supervised approaches to detect dynamic and adaptive fraud has focused the interest on unsupervised methods, which will be addressed next.

3.3.2 Unsupervised Algorithms

Due to the lack of labelled medical data and the shortcomings of supervised methods, unsupervised methods were developed. They primarily categorize claims and identify those that deviate from typical patterns. Since they do not require pre-labelled data, unsupervised methods may serve as initial filters that list the potentially fraudulent claims before the audit. This way, personnel costs decrease as fewer transactions are reviewed (Laleh and Azgomi, 2009). Unsupervised algorithms also have the advantage of being independent of a particular classified data set, allowing them to detect changing fraud patterns.

Even basic unsupervised approaches may be beneficial when combined with expertise regarding discriminating features (Copeland et al., 2012). Given the nature of unlabeled medical data, unsupervised learning is still seen as a valuable and promising technique, despite the necessity for extra review by subject matter experts (Bauder et al., 2017).

Lin et al. (2008) used clustering for the first time in medical data to segment general practitioner practice patterns. Then, Musal (2010) and Liu and Vasarhelyi (2013) use geolocation data within a clustering-based approach. The Bayesian Bernoulli co-clustering technique of Ekin et al. (2013) model's dyadic data concentrates on the occurrence of visits among providers and beneficiaries. The above statement can potentially reveal an emerging type of fraud called "conspiracy fraud" that involves attributes of more than one party in the medical system. These clustering algorithms help the investigator group the billings and variables of interest.

In the detection of medical fraud, outlier detection methods are widely used. Outliers correspond to observations outside the data's primary grouping and unexpected observations. A basic technique to detect outliers would be to rank data concerning variables of interest and label outcomes lower or higher than a predetermined threshold as outliers. Such a threshold can be determined using the knowledge of the entire data set or as a certain deviation measure (standard deviation) away from a central measure (mean, median). Capelleveen (2013) provides an overview of outlier detection methods and a variety of tests to assess their efficacy. Linear models, boxplots, peak analysis, multivariate clustering, and expert evaluation are outlier analysis methods. These outlier analysis methods include linear models, boxplots, multivariate clustering, peak analysis, and expert evaluation. Shan et al. (2009) propose a local density-based outlier identification approach for detecting improper billing patterns in Medicare in Australia. Ng et al. (2010) model Australia's Medicare Spatio-temporal data within an unsupervised anomaly detection framework. Lu and Boritz (2005) use Benford's Law Distributions to detect anomalies in claim reimbursements.

Tang et al. (2011) describe an integrated approach that combines feature selection, clustering, pattern recognition, and outlier detection to detect fraud in Australia's Medicare system. Carvalho et al. (2017) describe a two-phase anomaly detection method for detecting fraudulent hospitals in Brazil's public healthcare system. Outlier detection studies using prescription data are also available. A distance-based unsupervised algorithm for evaluating prescription fraud risk was described by Aral et al. (2012). Iyengar et al. (2014) constructed a normalized baseline behavioral model to find the abnormalities for each prescription area.

A case study for Medicaid dental practice investigations is presented by Van Capelleveen et al. (2016). They explore the implementation of several outlier detection methods utilizing different metrics. Bauder and Khoshgoftaar (2016) discuss outlier detection via Bayesian inference, which employs probability distributions and credibility intervals to examine outliers. Johnson and Nagarur's (2016) multi-stage methodology is also based on quantifying risk and distance from thresholds. Finally, Ekin et

al. (2017b) show how the concentration function can be used as a prescreening outlier detection tool to aid in detecting medical fraud.

In addition, industry tools based on graph analytics, association, and link analysis may allow investigators to identify relationships, links, and hidden patterns of information sharing and interactions within potentially fraudulent groupings of providers and patients. The number and quality of the links between businesses can be analyzed using the similarities in their contact information, locations, service providers, assets, and associates. Potential relationships with players that are found to be involved in fraud may provide red flags and lead to prospective investigations. These can help reveal organized, sophisticated, and collusive networks of providers and patients.

Unsupervised approaches are generally used to flag potentially fraudulent activities before bringing domain experts into the investigation. Therefore, close cooperation between physicians, statisticians, and people involved in decision-making would be beneficial during defining and tuning the model and analyzing and interpreting the results.

3.3.3 Hybrid Algorithms

Hybrid techniques aim to increase the efficacy of medical fraud detection by combining unsupervised and supervised approaches. For instance, unsupervised methods can choose the number of classes used in the classification process (He et al., 1997). The distance metric is optimized using a genetic algorithm in their usage of the k-nearest neighbour algorithm in detecting different types of fraud. Williams and Huang (1997) utilize clustering methods and label these clusters with a classification system. Clustering is proven to overcome the shortcomings of decision trees with larger data sets and multiple categories.

The use of medical information may increase classification performance at the expense of the cost and complexity of such models. For example, Yang and Hwang (2006) employ a pattern discovery to determine expected behavior before utilizing an outlier identification method. Rule extraction and outlier detection to compare provider characteristics was used by Major and Riedinger (2002). According to the study mentioned above and discussion, supervised approaches have attracted the most research effort.

According to the study mentioned above and the discussion

, supervised approaches have garnered the most research effort. However, supervised methods can only be applied when labelled fraud cases are available. At the same time, these kinds of labelled data are not always easy to collect. Moreover, due to subjectivity, labelled data is not remarkably accurate sometimes. Hence, unsupervised methods will be more applicable considering data availability and accuracy. As a result, more research into unsupervised fraud detection systems for health care data should be launched.

3.4 Case Studies

Many papers and research on fraud discovery using data analytics technologies were reviewed and discussed. Several investigators or academicians have devoted considerable work to exploring and investigating fraud detection models using data analytics. Below is some of the research on fraud detection in the health sector.

Richard and Taghil (2018) propose a machine learning approach for Medicare fraud detection using publicly available claims data and labels for known fraudulent medical providers. They successfully demonstrated the effectiveness of applying machine learning with random under-sampling to identify Medicare fraud. They employed random under-sampling building four class arrangements. The results revealed that C4.5 decision tree and logistic regression learners give the most fraud detection capability, especially for the 80:20 class arrangements, giving average AUC (Area Under ROC curve) scores of 0.883 and 0.882, respectively, with low wrong negative rates.

Inês (2017) created an artificial neural network that learns from insurance data and evolves continuously over time, anticipating fraudulent behaviors or actors and contributing to institutions' risk protection strategies.

Dharani and Shoba (2015) suggested a data analytics technique or approach for discovering fraudulent prescriptions in extensive prescription databases. They designed and built a personalized data analytics model for detecting prescription fraud. They utilized data analytics techniques for allocating a risk score to prescriptions about prescribed medicament-diagnosis uniformity, prescribed medicament uniformity within a prescription, prescribed medicament age, sex uniformity, and diagnosis-cost uniformity. The suggested model functions substantially well for the prescription fraud detection hitch, with a 77.4 per cent actual positive rate.

Pal and Pal (2015) suggested using varied data analytics technologies like ID3, J48, and Naïve Bayes to discover healthcare fraud. According to the outcomes, ID3 has the highest accuracy of 100 per cent, with J48 having the lowest accuracy of 96.7213 per cent.

Joudaki et al. (2015) implemented a data analytics methodology on a sizable health insurance institution dataset of non-governmental general physicians' prescription claims. Thirteen pointers were created in total. More than half, about 54 per cent, of the general physicians were culprits of performing remorseless behavior. The outcomes also determined that 2 per cent of physicians as fraud culprits. Discriminant analysis indicated that the indicators showed satisfactory effectiveness in discovering physicians who were fraud culprits, about 98 per cent and abuse, about 85 per cent, in a fresh data instance.

Liu (2014) has suggested a geolocation clustering design that analyses the geolocation particulars of Medicare and Medicaid recipients and service givers to identify doubtful claims with the rationale that recipients like to opt for health service providers somewhat shorter distances from where they live.

Shin et al. (2012) discovered misconduct in internal medicine outpatient clinics' claims by a risk score for showing the extent of the possibility of misconduct by healthcare givers and then grouped the healthcare givers with the use of a decision tree. They used a specific interpretation of the outlier score and obtained 38 features for identifying misconduct and corruption.

3.5 Chapter Summary

Chapter 3, focused on medical claims data, explains why they have a complex nature, and a thorough search in the literature for the applications of data analytics and statistical machine learning that have been used along with case studies.

CHAPTER 4

Machine Learning Techniques used in Fraud Detection

4.1 Healthcare Fraud Detection Methods

Nowadays, fraud detection uses advanced statistical and machine learning methods to struggle with fraud. Another commonly used method is data analytics, which attempts to understand data and identify patterns using a statistical method. Data analytics is a multidisciplinary field that combines not only statistics and machine learning techniques but also visualization, computer science, and database technology.

There are different elaborate and complicated patterns with many small trivial requirements incorporated into fraud whose data is accumulated over a protracted time frame. It is challenging to detect these patterns when we have a vast accumulation of data and few means for evaluating them. Traditionally, a few auditors used to handle thousands of health care claims. Thus, usually, only experienced investigators manage fraud detection. However, the extensive data collection makes this method time-consuming and inefficient. Improvements in data analytics and machine learning tools bring attention to automated systems for fraud detection. For the detection of anomalies and fraud detection, behavioral profiling methods based on machine learning techniques are used, and for this purpose, the behavior pattern of each person involved in the healthcare system is configured to observe and check for any deviation from the standards.

The methods for detecting fraud and corruption in healthcare can be classified into three different methods. Supervised methods are used when historical data is available; unsupervised methods are used when there is no historical data, and a hybrid method combines both supervised and unsupervised methods. Supervised healthcare methods frequently use specific healthcare rules and statistical studies to classify the situations. On the other hand, the unsupervised group's case has similar characteristics.

Table 1 presents a resume of the different fraud detection methods, distinguishing the type of detection, the method of detection, and the fraudulent behavior in the case. After the methods to detect fraud and corruption are presented, three technical adjustments are made, improving the algorithm's efficiency by interpreting the healthcare situation and adapting the models to its needs.

Supervised Methods	**Neural Networks**
	-Service Providers Fraud
	-Insurance Subscribers' Fraud
	Decision Tree
	- Service Providers Fraud
	- Insurance Subscribers' Fraud
	Genetic Algorithm and KNN
	-Service Providers Fraud
	-Rule-based Classifier and BN
	-Insurance Subscribers' Fraud Unsupervised Methods
Unsupervised Methods	**SOM**
	-Service Providers Fraud
	Association Rules
	-Insurance Subscribers' Fraud
	-Rule-based Method
	-Service Providers Fraud
	-Finite Mixture Model
	-Insurance Subscribers' Fraud
	Clustering
	-Service Providers Fraud
	-Subjective Utility Model
	-Insurance Subscribers' Fraud Hybrid Methods
Hybrid Methods	**SOM and Neural Networks**
	-Service Providers Fraud
	-Clustering and Decision Tree
	-Insurance Subscribers' Fraud

Table 1: Resume of Healthcare Fraud Detection Methods

Several supervised methods have been used in health care fraud detection, including Neural Networks (NNs), decision trees, fuzzy logic, and Bayesian networks. The two most popular methods are NNs and decision trees, as evidenced in Figure 2. In applying supervised methods to health care fraud detection, there is an increasing trend to combine several supervised methods to improve classification performance.

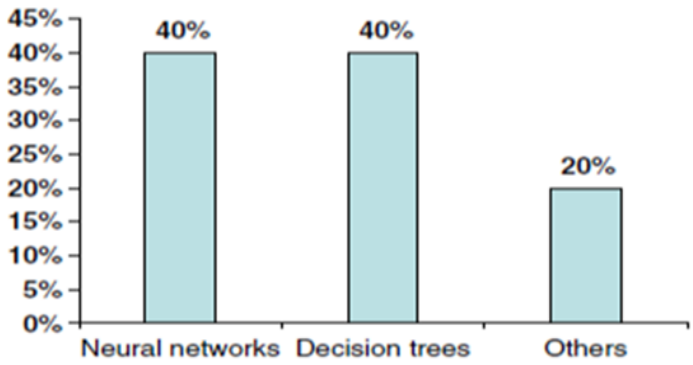

Figure 2: Percentages of Papers on Different Supervised Methods

4.2 Supervised Methods

4.2.1 Decision Tree

4.2.1.1 What a Decision Tree is

A decision tree can be described as a map of the possible outcomes of a series of related choices. It enables an individual or organization to compare various actions based on their costs, probabilities, and benefits. They can be used to spark informal conversation or to sketch an algorithm that mathematically predicts the optimal option.

A decision tree usually begins with a single node and branches into possible outcomes. Each of those outcomes leads to other nodes, branching into other possibilities. So, this gives it a treelike shape. Nodes can be classified into chance nodes, decision nodes, and end nodes. A chance node, represented by a circle, displays the Probability of specific results. A decision node, represented by a square, displays a decision to be made, while an end node, represented by a triangle, shows the outcome of a decision path. Figure 3 is represented a simple example of a decision tree.

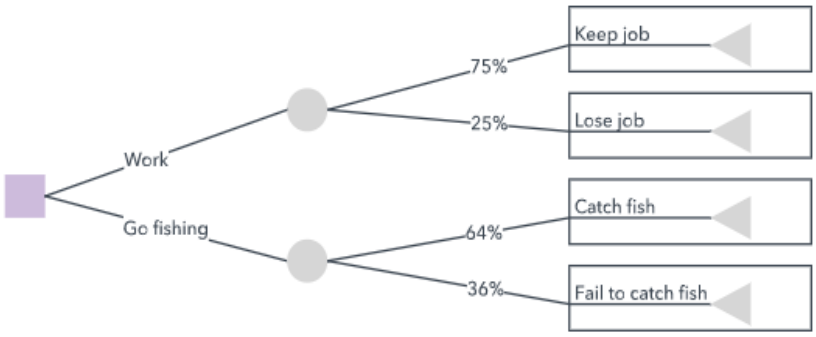

Figure 3: Example of Decision Tree

Flowchart symbols can also be used to design decision trees, which some people find easier to read and interpret. Figure 4 is represented the symbols of decision trees.

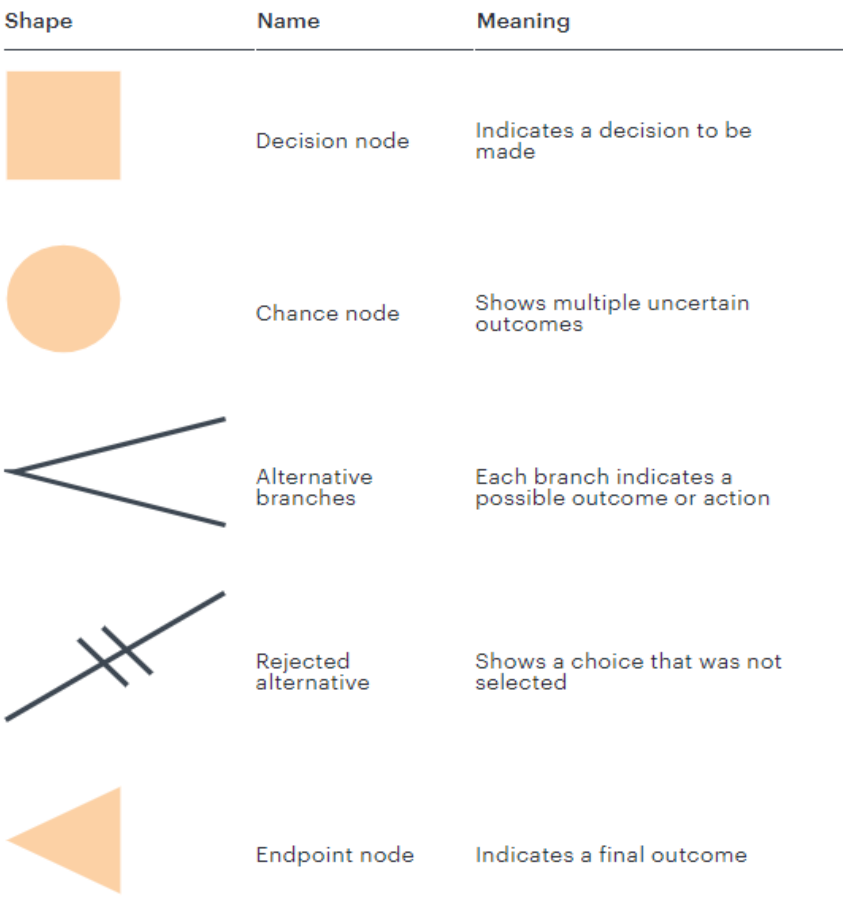

Figure 4: Decision Tree Symbols

4.2.1.2 Advantages and Disadvantages of Decision Trees

There are many advantages to decision trees, and a few of them are listed below.

- The cost of using the tree to predict data decreases with each additional data point.
- It can be used with categorical or numerical data.
- Can model problems with multiple outputs.
- It makes results easy to explain, as it uses a white-box model.
- The reliability of a tree can be measured and tested.
- Its tendency to be accurate regardless of whether it violates the assumptions of source data.
- How easy they are to understand.
- They can be helpful with or without complex data, and any data requires minimal preprocessing.
- Existing trees can be expanded with new options.
- They are worth determining the best option out of several.
- The simplicity with which they can be combined with other decision-making tools.

However, they also have a few disadvantages:

- When working with categorical data that has multiple levels, the information gained is biased toward the attributes with the most levels.

- Calculations can grow complex when dealing with uncertainty and lots of related outcomes. Decision trees can become excessively complex. A more compact influence diagram can be a suitable alternative in such instances. Influence diagrams reduce the focus on critical decisions, inputs, and objectives.
- Conjunctions between nodes are limited to AND, but decision graphs allow for nodes linked by OR (Odds Ratio).

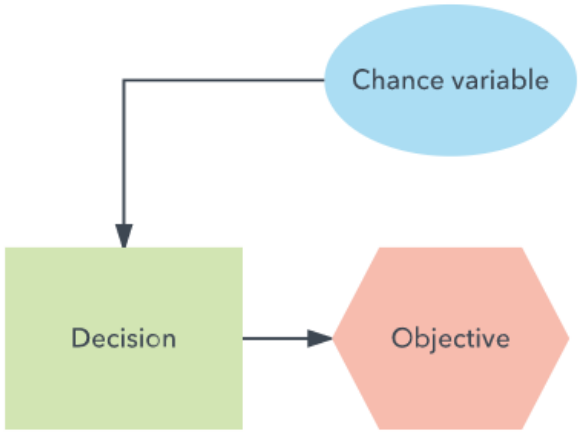

Figure 5: Influence Diagram

4.2.1.3 Decision Trees in Machine Learning and Data Analytics

A decision tree can also be used to help develop automated predictive models, which have applications in machine learning, data analytics, and statistics. Known as decision tree learning, this method considers observations about an item to predict that item's value.

In these decision trees, nodes reflect data rather than decisions, and this type is also known as a classification tree. Each branch comprises a set of attributes, or classification rules, associated with a particular class label found at the branch's end. These rules, also known as decision rules, can be represented in an if-then clause, with each decision or data value forming a clause, such that, for instance, "If conditions 1, 2, and 3 are fulfilled, then outcome x will occur with y certainty."

Each new piece of data helps the model predict which of a finite set of values the subject belongs to. A bigger decision-making model includes all these data. The actual number, such as a price, is sometimes the predicted variable. Decision trees with continuous, infinite possible outcomes are termed regression trees.

For improved accuracy, occasionally, multiple trees are used together in ensemble methods:

- Bagging generates multiple trees by resampling the source data and then polls those trees to reach a consensus.
- A Random Forest classifier consists of multiple trees designed to improve classification rates.
- The boosted trees can be used for regression and classification trees.

- The trees in a Rotation Forest are all trained by utilizing PCA (Principal Component Analysis) on a random chunk of the data.

A decision tree is ideal when it represents the most critical data with the fewest levels or questions. Building association rules and setting the target variable on the right can also be used to create a decision tree. Each method must identify the best way to split the data at each level. Standard methods include assessing the Gini impurity, information gain, and variance reduction.

4.2.1.4 Decision Tree Uses

The decision trees are used to identify service providers' fraud, detect insurance subscribers' fraud, and plan audit strategies for fraud detection. Decision trees map include information by several classifiers following politics, such as minimizing false positives, minimizing false negatives, and achieving a tradeoff between false positives and false negatives. Minimizing false positives minimizes wasteful costs for unnecessary fraud investigation, and minimizing false negatives maximizes detection ability.

On the other hand, Decision Tree also uses adaptive boosting to reduce misclassifications errors and produces a set of classifiers and votes on them to classify cases. The various classifiers are constructed sequentially, focusing on training cases that prior classifiers have misclassified.

It also applied a divide-and-conquer technique to detect insurance subscribers' fraud. First, all insurance subscribers' profiles are divided into groups. Then, a decision tree is built for each group and converted to a set of rules. Each rule is evaluated according to its significance through statistics at the end.

Decision trees are an excellent approach to detecting fraud in healthcare because of their simplicity in interpreting the results and ability to generate rules from the tree and handle missing values. However, too many rules were generated for large dimensional databases, and few adjustable parameters were available.

4.2.2 Logistic Regression

4.2.2.1 What Logistic Regression is

This type of statistical analysis is often used for predictive analytics and modelling and extends to applications in machine learning. In this analytics approach, the dependent variable is categorical or ordinal: A or B, a binary regression, or a range of finite options A, B, C, or D, a multinomial regression. Statistical software uses statistical software to understand the relationship between the dependent variable and one or more independent variables by estimating probabilities using a logistic regression equation.

Christopher M. Bishop, in his book Pattern Recognition and Machine Learning, defines the logistic sigmoid function $\sigma(a)$ as below:

$$\sigma(\alpha) = \frac{1}{1+exp(-\alpha)} \tag{4.1}$$

Which is plotted in Figure 6. The term 'sigmoid' means S-shaped. This function is sometimes called a 'squashing function' because it maps the whole real axis into a finite interval. The logistic sigmoid plays an essential role in many classification algorithms. It satisfies the following symmetry property

$$\sigma(-\alpha) = 1 - \sigma(\alpha) \tag{4.2}$$

as is easily verified. The inverse of the logistic sigmoid is given by

$$\alpha = \ln\left(\frac{\sigma}{1-\sigma}\right) \tag{4.3}$$

Moreover, it is known as the logit function.

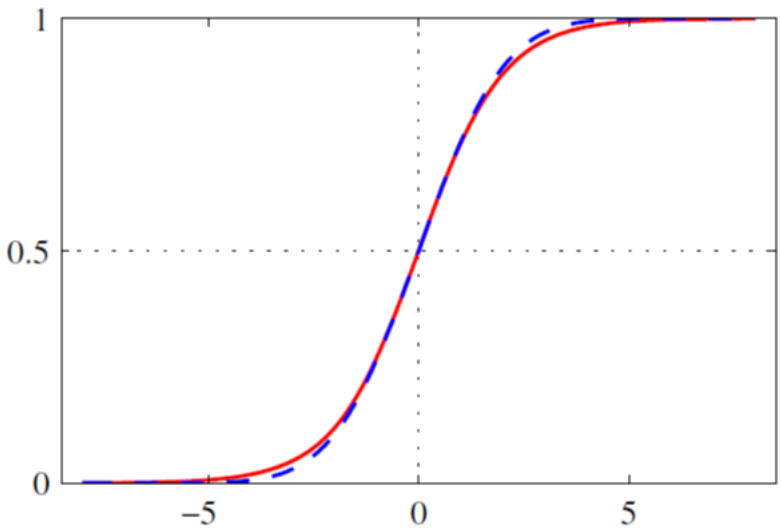

Figure 6: Plot of the logistic sigmoid function σ(a)

This analysis can help predict the likelihood of an event or a choice being made. For example, when someone wants to know the likelihood of a visitor choosing an offer made on a website or not, this is the dependent variable. The analysis can look at known characteristics of visitors, such as sites they came from, repeat visits to the site, and behavior on the site; these are the independent variables. Logistic regression models help to determine the Probability of what type of visitors are likely to accept the offer or not. As a result, better decisions can be made about promoting the offer or decisions about the offer itself.

4.2.2.2 Logistic Regression in Machine Learning and Data Analytics

Machine learning uses statistical concepts to enable machines, or computers, to "learn" without explicit programming. A logistic approach fits best when the machine learns the task is based on two values or a binary classification. A simple example is a computer that could use this type of analysis to make determinations about promoting the offer and take actions all by itself. Moreover, as more data is provided, it could learn how to do this better over time.

Some types of predictive models that use logistic analysis:

- Generalized linear model
- Discrete choice
- Multinomial logit
- Mixed logit
- Probit
- Multinomial probit
- Ordered logit.

Certainly, multinomial analysis can help to examine a range of categorical outcomes: A, B, C, or D. But binary analysis, yes or no, present or absent, is more often used. Although the outcomes are constrained, the possibilities are not. Binary logistic regression can examine everything from baseball statistics to landslide susceptibility to handwriting analysis.

This approach to analytics also proves useful for a range of statistical concepts and applications:

- Text analytics
- Chi-square automatic interaction detection (CHAID)
- Conjoint analysis
- Bootstrapping statistics
- Nonlinear regression
- Cluster statistics and cluster analysis software
- Monte Carlo simulation
- Descriptive statistics

Statistical analysis software delivers excellent value for approaches such as logistic regression, multivariate analysis, neural networks, decision trees, and linear regression. Nevertheless, remember that hardware and cloud-computing solutions should also be considered if there is a need to accommodate large data sets either on-premises, in the cloud, or in a hybrid cloud configuration.

4.2.2.3 Assumptions and Limitations of Logistic Regression

While binary logistic regression is more generally used and discussed, it can be helpful to evaluate when each form is most effective.

Multinomials can classify subjects based on a categorical range of variables for behavior prediction. For instance, conduction of a survey in which participants are asked to identify one of several competing products as their favourite. The creation of profiles of people most likely to be interested in the product and plan the advertising strategy accordingly.

Binary is most useful when there is a need to model the event probability for a categorical response variable with two outcomes. A loan officer needs to determine whether the next customer is likely to

default or not default on a loan. Binary analysis can help estimate the risk of extending credit to a particular customer.

Understanding when this type of analysis could be ineffective is also helpful. Below are some hazards to be aware for:

- Independent variables must be valid; incorrect or incomplete variables will degrade a model's predictive value.
- Avoid continuous outcomes. Temperatures, time, or anything open-ended will make the model much less precise.
- Do not use inter-related data. The model will tend to outweigh the significance of some observations if they are related.
- Be wary of overfitting or overstatement. These statistical analysis models are precise, but the accuracy is not infallible or without variance.
- When to use linear or logistic analysis is a common query. Linear regression analysis is more effectively applied when the dependent variable is open-ended or continuous, astronomical distances or temperatures, for example. Use the logistic approach when the dependent variable is limited to a range of values or categorical, A or B...or A, B, C, or D.

4.2.2.4 Logistic Regression Uses

Predictive models built using this approach can positively impact a business or organization. Because these models help to understand relationships and predict outcomes, they are used to improve decision-making. For example, a manufacturer's analytics team can use logistic regression analysis as part of a statistics software package to estimate the Probability between part failures in machines and the length of time those parts are held in inventory. With the information it receives from this analysis, the team can decide to adjust delivery schedules or installation times to eliminate future failures.

In medicine, this analytics approach can predict the likelihood of disease or illness for a given population, which means that preventative care can be implemented. Businesses can use this approach to uncover patterns that lead to higher employee retention or create more profitable products by analyzing buyer behavior. In the business world, this type of analysis is applied by data scientists whose goal is clear: to analyze and interpret complex digital data.

In healthcare, the logistic regressions used to detect fraud are binary logistic regressions, where the classification variable can be true/false, for a model, or success/failure, for a treatment. Logistic regression measures the relationship between the dependent and independent variables by estimating probabilities using a logistic function.

It has defined a probability that a given institution is operating lawfully and can be defined as the “odds ratio” for a fraudulent and non-fraudulent institution. The logistic regression function has the advantage of being easily interpreted.

4.2.3 Bayesian Network

4.2.3.1 What a Bayesian Network is

The formal definition of a Bayesian network consists of a graphical model (namely, the directed acyclic graph) together with the corresponding probability potentials, as Timo Koski and John M. Noble stated in their book "Bayesian Network: An introduction" (2009). In other words, a Bayesian network is a probabilistic graphical modelling technique that employs the concept of Probability to calculate uncertainty. Using directed acyclic graphs are used to model improbability.

The Bayesian network concept is based on Bayes' theorem, which allows us to represent the conditional probability distribution of cause given empirical evidence using the converse conditional probability distribution of observing evidence given cause. For estimating class-conditional densities for subsequent use in Bayes' theorem to find posterior probabilities. In most practical pattern classification problems, using more than one feature variable is necessary. We may also wish to consider more than two possible classes so that we might consider more than two characters in our character recognition problem. For c different classes C_i,..., C_c, and for a continuous feature vector x, we can write Bayes' theorem in the form, which can now be written in the form as follows (Timo Koski and Jonh M. Noble, 2009):

$$P(C_k|x) = \frac{p(x|C_k)P(C_k)}{p(x)} \tag{4.4}$$

where the unconditional density p(x) is given by

$$p(x) = \sum_{k=1}^{c} p(x|C_k)P(C_k) \tag{4.5}$$

which ensures that the posterior probabilities sum to unity

$$\sum_{k=1}^{c} P(C_k|x) = 1 \tag{4.6}$$

In practice, we might choose to model the class-conditional densities $p\left(x\middle|C_{f_c}\right)$ by parametrized functional forms. When viewed as functions of the parameters, they are referred to as likelihood functions for the observed value of x. Bayes' theorem can therefore be summarized in the form

$$posterior = \frac{likelihood\ x\ prior}{normalization\ factor} \tag{4.7}$$

4.2.3.2 Bayesian Network Uses

Bayesian networks provide a straightforward mathematical language to express relations between variables transparently. In many engineering examples, the variables that should be present in the model are well defined. From an appropriate model that contains the hidden (or non-observable) variables and the observable variables, and where it is clear which variables may be intervened on, it will be possible to verify whether certain 'identifiability' conditions hold and hence to conclude whether or not there is a causal relation from the data, without a controlled experiment.

The Bayesian approach provides the framework for both assessing uncertainties regarding fraudulent behavior as well as for making decisions for the investigation of fraud. The Naïve Bayesian classifier is a precise chancy classifier centred on implementing the Bayesian theorem (from Bayesian statistics) with firm (naïve) objectivity expectations. The Naïve Bayes algorithm uses the Bayes theorem for its classification. This is by calculating the possibilities of the feature values for each classification group and using the possibilities to guess the class of the undefined instances.

The Naïve Bayes classifier method is specifically adapted if the dimension of entered data is higher. Naïve Bayesian classifier supposes that the elements or attributes characters are hypothetically independent and there consist no dependence associations amongst the attributes. This makes Naïve Bayes the most accurate classifier when the presumption holds. The merit of the Naïve Bayes classifier is that it only needs a limited quantity of the training data to compute the means and variances of the parameters required for classification. Since independent parameters are hypothesized, only the variances of the parameters within all the labels require to be affected and not the whole covariance matrix.

4.2.4 Neural Networks

4.2.4.1 What Neural Networks are

Neural networks, also known as Artificial Neural Networks (ANNs) or simulated neural networks (SNNs), are a subset of machine learning that provide the foundation of deep learning algorithms. Their name and structure are inspired by the human brain, replicating how biological neurons signal to one another. The brain is a highly complicated, nonlinear, parallel computer (information processing system). It can organize its structural parts, known as neurons, to do specific calculations (e.g., pattern recognition, perception, and motor control) many times faster than the fastest digital computer. Consider human eyesight, which is an example of an information-processing task. The visual system's role is to produce an image of the environment around us and, more importantly, to offer the information we need to interact with the environment. More specifically, the brain frequently completes perceptual identification tests (such as recognizing a familiar face embedded in an unfamiliar scene) in 100-200 ms, but tasks of far lower complexity take much longer on a powerful computer.

Plasticity allows the developing nervous system to adapt to its surroundings. Plasticity appears to be fundamental to the operation of neurons as information-processing units in the human brain and also appears to be crucial to the functioning of neural networks made up of artificial neurons. In its most general form, a neural network is a machine designed to model how the brain performs a particular activity or function of interest; the network is often constructed using electronic components or is simulated in software on a digital computer. Neural networks use a massive interconnection of simple computing cells known as "neurons" or "processing units" to attain high performance. Below is a

definition of a neural network viewed as an adaptive machine offered by Simon Haykin in his book Neural Networks and Learning Machines (2009):

A neural network is a massively parallel distributed processor made up of simple processing units with a natural propensity to store experiential knowledge and make it available for use. It resembles the brain in two respects:

1. *The network's environment acquires knowledge through a learning process.*
2. *Interneuron connection strengths, known as synaptic weights, are used to store the acquired knowledge.*

A learning algorithm is a procedure used to accomplish the learning process, and its goal is to adjust the synaptic weights of the network systematically to achieve the desired design objective.

The usual way of designing neural networks is to change synaptic weights. This technique is the most similar to linear adaptive filter theory, which is already well known and widely used in various applications (Widrow and Stearns, 1985; Haykin, 2002). However, a neural network may change its topology, which is prompted by the fact that neurons in the human brain can die, and new synaptic connections can grow.

Artificial Neural Networks are formed of a node layer, incorporating an input layer, one or more hidden layers, and an output layer. Each node, or artificial neuron, links to another and has a corresponding weight and threshold. If the output of any individual node is above the given threshold value, that node is activated, transmitting data to the next layer of the network. Otherwise, no data is transmitted to the next network layer. In Figure 7, a deep neural network is represented.

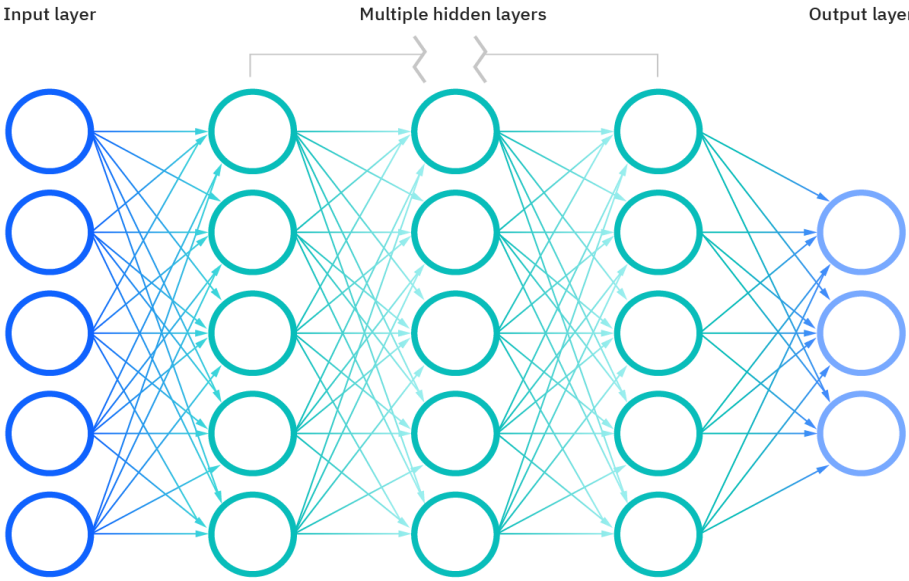

Figure 7: Deep Neural Network

Deep Learning and neural networks tend to be used interchangeably in conversations, which can be confusing. As a result, it is crucial to note that the "deep" in deep learning only refers to the depth of layers in a neural network. A neural network that comprises more than three layers, including the inputs

and the output, can be regarded as a deep learning algorithm. A neural network that includes two or three layers is only a primary neural network.

Neural networks rely on training data to learn and increase their accuracy over time. However, once these learning algorithms are fine-tuned for accuracy, they are vital tools in computer science and artificial intelligence, allowing us to classify and cluster data at a high velocity. Compared to manual identification by human experts, speech recognition or image recognition tasks can take minutes rather than hours. Google's search algorithm is one of the most well-known neural networks.

4.2.4.2 Models of a Neuron

A neuron is a basic information-processing unit in the operation of a neural network. The block diagram in Figure 8 illustrates a neuron model, which serves as the foundation for creating a broad family of neural networks. We identify three fundamental features of the neural model:

1. A set of synapses or connecting links, each of which has its weight or strength. In particular, the synaptic weight w_{kj} is multiplied by a signal x_j at the input of synapse j linked to neuron k. It is critical to pay attention to how the subscripts of the synaptic weight wkj are written. The first subscript in w_{kj} indicates the neuron in question, while the second indicates the input end of the synapse to which the weight refers. Unlike the weight of a synapse in the brain, the synaptic weight of an artificial neuron can have both negative and positive values.

2. An adder for summing the input signals, weighted by the neuron's synaptic strengths; the processes described here create a linear combiner.

3. An activation function for restricting the amplitude of a neuron's output.

The activation function is also known as a squashing function since it squashes the output signal is allowed amplitude range to some finite value.

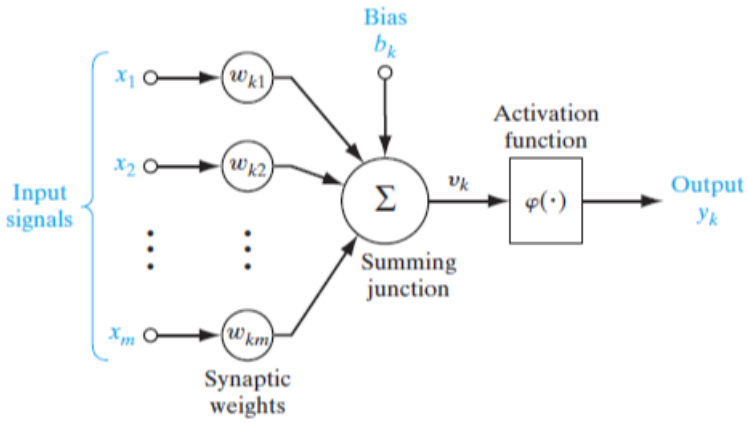

Figure 8: Nonlinear model of a neuron, labelled k.

Typically, the closed unit interval [0,1] or [-1,1] is used to represent the normalized amplitude range of a neuron's output. An external bias, represented by b_k, is also included in the neural model of Figure 8. Whether the bias b_k is positive or negative, it can either increase or decrease the net input of the

activation function. The following two equations offered in Haykin's book (2009), which were previously referred to and may be used to represent the neuron k shown in Figure 11 mathematically:

$$u_k = \sum_{j=1}^{m} w_{kj} * x_j \tag{4.8}$$

and

$$y_k = \varphi(u_k + b_k) \tag{4.9}$$

where x_1, x_2, ..., x_m are the input signals; $w_{k1}, w_{k2}, \ldots, w_{km}$ are the respective synaptic weights of neuron k; u_k (not shown in Figure 8) is the linear combiner output due to the input signals; b_k is the bias; φ(·) is the activation function, and y_k is the output signal of the neuron. The use of bias b_k has the effect of applying an affine transformation to the output u_k of the linear combiner in the model of Figure 8, as shown by

$$v_k = u_k + b_k \tag{4.10}$$

As we think about more practical use cases for neural networks, such as image recognition or classification, we will employ supervised learning, or labelled datasets, to train the algorithm. We will use a cost or loss, function to evaluate the model's accuracy as we train it. The mean squared error (MSE) is another term for this. In the equation below,

i represents the index of the sample,

y-hat is the predicted outcome,

y is the actual value, and

m is the number of samples.

$$Cost\ Function = MSE = \frac{1}{2m}\sum_{i=1}^{m}(\hat{y} - y)^2 \tag{4.11}$$

Ultimately, we aim to reduce our cost function to ensure fit validity for every observation. As the model adjusts its weights and bias, it employs the cost function and reinforcement learning to reach the point of convergence or the local minimum. The algorithm modifies its weights by gradient descent, allowing the model to determine the direction to pursue to reduce errors or minimize the cost function. With each training sample, the model's parameters progressively adjust to converge at the minimum. In Figure 9, there is a representation of minimizing the cost function, as referred to before.

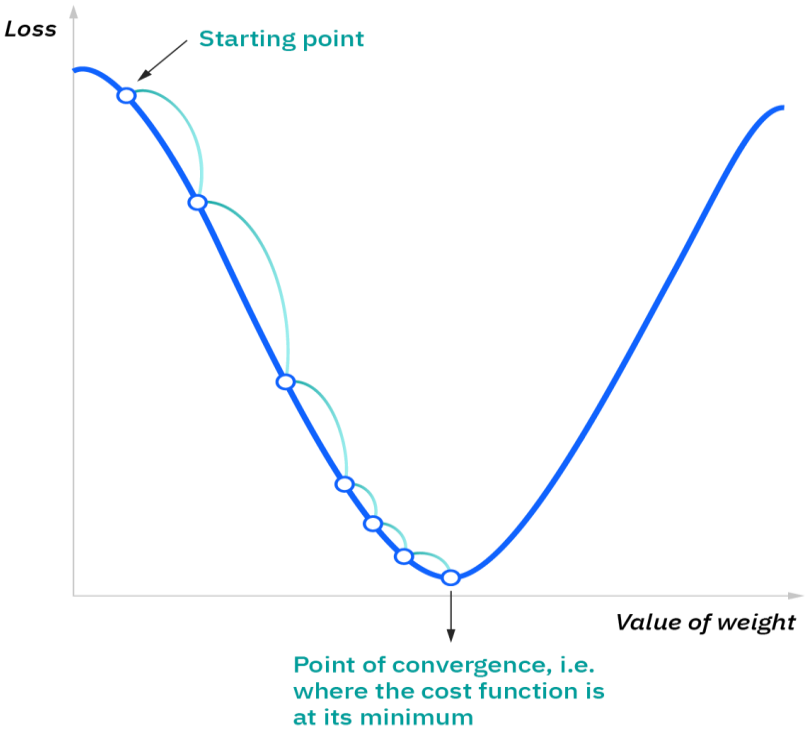

Figure 9: Representation of Minimizing the Cost Function

Most deep neural networks are feedforward, meaning they only flow in one direction, from input to output. However, you can also train your model by backpropagation, going oppositely from output to input. Backpropagation allows us to calculate and calculate the error associated with each neuron, allowing us to alter and match the parameters of the model(s) appropriately.

4.1.4.3 Types of Neural Networks

Neural networks can be categorized into different types, which are utilized for different purposes. While this is not a complete list of types, the below would be representative of the most popular kinds of neural networks that are existed across its everyday use cases:

The perceptron is the oldest neural network, constructed by Frank Rosenblatt in 1958. It has a single neuron and is the simplest form of a neural network, as Figure 10 shows.

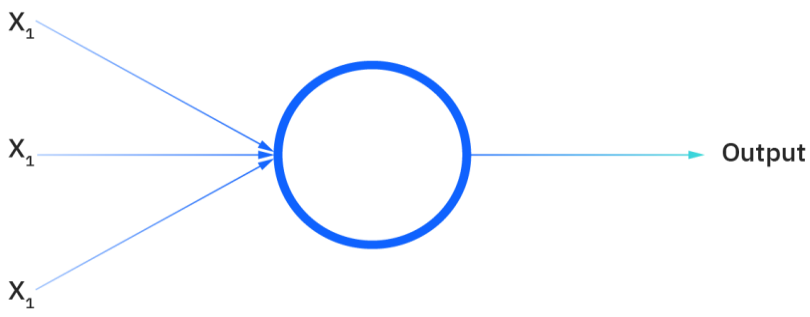

Figure 10: The Perceptron by Frank Rosenblatt

Feedforward neural networks, or multilayer perceptrons (MLPs), comprise an input layer, a hidden layer or layers, and an output layer. While these neural networks are frequently referred to as MLPs, it is essential to mention that they are comprised of sigmoid neurons, not perceptrons, as most real-world problems are nonlinear. Data usually is input into these models to train them, and they are the foundation for computer vision, natural language processing, and other neural networks.

Convolutional neural networks (CNNs) are similar to feedforward networks; however, they are commonly utilized for image recognition, pattern identification, and computer vision. These networks harness principles from linear algebra, notably matrix multiplication, to find patterns inside an image.

Recurrent neural networks, also known as RNNs, are identified by feedback loops. These learning algorithms are generally used when analyzing time-series data to create predictions about future outcomes, such as stock market predictions or sales forecasting.

4.2.4.4 Advantages and Disadvantages of Neural Networks

Under the machine-learning methods, there are Artificial Neural Networks. It is inspired by the animals' biological nervous system, the brain, when it processes information. It is mainly composed of highly connected elements, like neurons, that work together to solve the problem. Neural Networks are especially useful because they handle significant amounts of data and complex structures with no linear relationships.

Some nodes receive scalar input from other nodes and transform the information into a single output signal. The interconnections are weighted, and the weights are tunable. This method helps detect fraud in healthcare due to its enormous variable and quantity of data. This method also tolerates noisy data.

Neural networks classify practice profiles of general practitioners to reduce the inconsistency of experts' classifications due to subjectivity. Neural Networks typically use three layers for data classification: input, hidden, and output. In the hidden layer, nodes receive a weighted sum of the input variables and transform the sum into an output signal, considering a certain threshold. The output layer nodes convert the signal from the hidden layer into a classification signal. Then Neural Network establishes a relationship between the input and output data.

Besides all the advantages, the typical concerns on Neural Networks are overfitting and a small training sample size with many features. The first, overfitting, produces a small error on the training dataset but a much larger error when new inputs are added. The second leads to high variance between different runs in the Neural Network. Healthcare data is mainly sensitive to overfitting because there are considerably more legitimate cases than fraudulent ones.

One way to solve the overfitting problem is to add a weighted delay to the error function for training a Neural Network. This strategy intends to achieve a tradeoff between the training error and the complexity of the Neural Network. Another way to solve the overfitting problem is the early stopping

technique used for a different dataset. The first is to update the weights and biases points; the other is to stop training when the network begins to overfit data.

A commitment to individual training instead of only one in the Neural Network was applied to control the problem with the variance. On average, the variance decreases successfully from 8 per cent of variance to 1.8 per cent of the variance.

Neural Networks is a complex machine learning method; usually, it is hard to understand how it works. There are still few approaches to handle overfitting, and usually, there are too many parameters to be tuned.

4.2.4.5 Historical Evolution of Neural Network

The history of neural networks is more extended than most people assume. While the idea of "a machine that thinks" can be dated to the Ancient Greeks. Below are the critical events that led to the growth of thought around neural networks, whose popularity has ebbed and flowed over time:

1943: Warren S. McCulloch and Walter Pitts published "A logical calculus of the ideas immanent in nervous activity" This research tried to understand how the human brain might form complex patterns through connected brain cells or neurons. One of the primary ideas that came out of this work was the comparison of neurons with a binary threshold to Boolean logic (i.e., 0/1 or true/false statements).

1958: Frank Rosenblatt is credited with the development of the perceptron, described in his research, "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain". He expands on McCulloch and Pitt's work by incorporating weights into the equation. Leveraging an IBM 704, Rosenblatt got a computer to learn how to discriminate cards marked on the left vs those on the right.

1974: While other researchers contributed to the theory of backpropagation, Paul Werbos was the first person in the US to acknowledge its implementation within neural networks within his PhD thesis.

1989: Yann LeCun released a paper describing how the use of restrictions in backpropagation and its incorporation into the neural network design can be utilized to train algorithms.

4.2.5 Support Vector Machines

4.2.5.1 What Support Vector Machines are

Support Vector Machines (SVM) are supervised machine learning algorithms that can be used for classification and regression. SVMs are more commonly used in classification issues. SVMs are based on the principle of finding a hyperplane that best divides a dataset into two classes, as seen in the graphic below. Figure 11 is a representation of finding a hyperplane that best divides a dataset into two classes of Support Vector Machines.

Support vectors are the data points closest to the hyperplane, the focus of a data set that, if removed, would modify the position of the dividing hyperplane. Because of this, they might be regarded as the critical aspects of a data set.

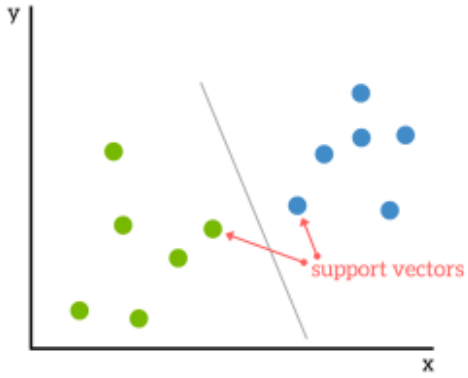

Figure 11: Representation of the Best Hyperplane of SVM

4.2.5.2 The hyperplane of SVM

As a simple example, for a classification task with only two features, as Figure 12 shows, we can think of a hyperplane as a line that linearly separates and classifies a data set.

Intuitively, the further our data points lie from the hyperplane, the more confident we are that they have been correctly classified. Hence, we need our data points to be as far away from the hyperplane as conceivable while still being on its right side. So, when additional testing data is added, whatever side of the hyperplane it lands will decide the class we give.

The margin is the distance between the hyperplane and the nearest data point from either set. The purpose is to choose a hyperplane with the most significant possible gap between the hyperplane and any point within the training set, giving a better likelihood of new data being categorized correctly.

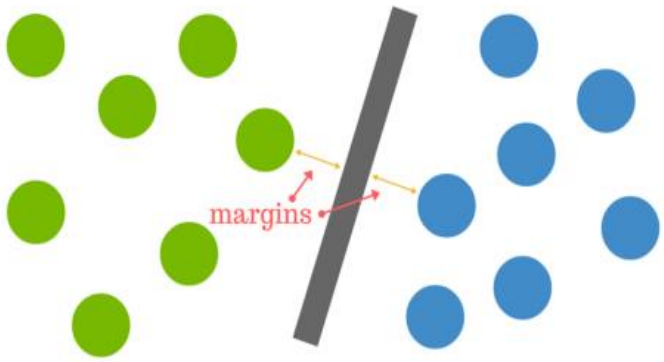

Figure 12: Representation of the Margin of SVM

This is where it can get tricky. Data is rarely ever as clean as our primary example above. A dataset will often appear more like the jumbled balls below, which show a linearly non-separable dataset. Figure 13 is a representation of a linearly non-separable dataset of SVM

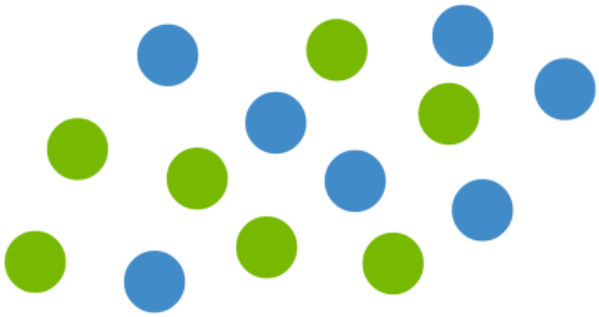

Figure 13: Representation of the Non-Separable Dataset of SVM.

Classifying a dataset like the one above requires moving from a 2-dimensional view of the data to a 3-dimensional one. It is easier to explain this using another simplified example. Assume our two sets of coloured balls are sitting on a sheet, abruptly lifted, launching the balls into the air. While the balls are in the air, you use the sheet to separate them. This 'lifting' of the balls signifies data mapping into a higher dimension. This is known as kernelling. Figure 14 displays the process of moving away from a 2d view of the data to a 3d view of an SVM.

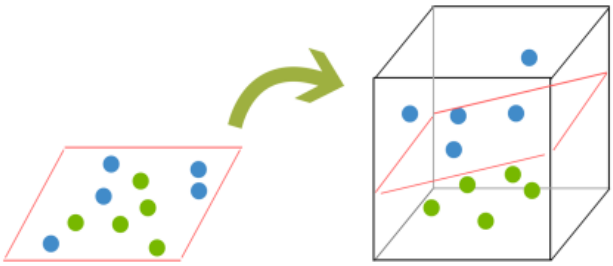

Figure 14: Process of the Movement from 2-Dimensional View to a 3-Dimensional of an SVM

Our hyperplane can no longer be a line since we are in three dimensions. It must now be a plane, as indicated in the example above. The concept is that the data will continue to be mapped into higher and higher dimensions until a hyperplane can be built to segregate it.

4.2.5.3 Advantages and Disadvantages of SVM

Using SVM in machine learning has various advantages:

- Accuracy
- Works well on smaller, cleaner datasets
- Because it just uses a subset of training points, it is more efficient

However, they also have a few disadvantages:

- It is not suited to larger datasets as the training time with SVMs can be high
- On noisy datasets with overlapping classes, it performs relatively poorly.

4.2.5.4 SVM Uses

SVM is used for category assignment, text classification tasks, spam detection, and sentiment analysis. It is also often used for image recognition challenges, performing notably well in aspect-based and colour-based classification. SVM also plays a significant role in several fields of handwritten digit recognition, such as postal automation services.

This supervised method in machine learning analyses data to recognize patterns. Support Vector Machines is a Kernel method, a class of algorithms for pattern analysis, which represents the similarity between two objects representing then in a dot in a vector space. The method constructs a linear discriminating function that separates two classes from support vectors formed by selecting a small number of critical boundary instances.

SVM does not need a large training dataset, and the training converges to a unique global solution. SVM constructs a hyperplane in a high dimensional space, which can be used for classification and linear regression.

Some studies suggest that Support Vector Machines perform better than Neural Networks because it needs fewer parameters to achieve the same accuracy. SVM works with a larger, transformed version of the feature space and finds a maximum margin hyperplane that separates two classes of the data.

This supervised method in machine learning analyses data to recognize patterns. Support Vector Machines is a Kernel method, a class of algorithms for pattern analysis, which represents the similarity between two objects representing then in a dot in a vector space. The method constructs a linear discriminating function that separates two classes from support vectors formed by selecting a small number of critical boundary instances.

SVM does not need a large training dataset, and the training converges to a unique global solution. SVM constructs a hyperplane in a high dimensional space, which can be used for classification and linear regression.

Some studies suggest that Support Vector Machines perform better than Neural Networks because it needs fewer parameters to achieve the same accuracy. SVM works with a larger, transformed version of the feature space and finds a maximum margin hyperplane that separates two classes of the data.

4.3 Unsupervised Methods

4.3.1 Principal Component Analysis

4.3.1.1 What Principal Component Analysis is

PCA is a dimensionality reduction method typically used to reduce the dimensionality of large data sets by reducing a large set of variables into a smaller one that still maintains much of the information in the large set.

Reducing the number of variables in a data set naturally comes at the expense of accuracy, but the idea of dimensionality reduction is to exchange a little accuracy for simplicity. Because smaller data sets are easier to examine and visualize, they make analyzing data easier and faster for machine learning algorithms without superfluous variables.

Statistically, PCA finds lines, planes, and hyper-planes in the K-dimensional space that approximate the data as well as possible in the least-squares sense. The variance of the coordinates on a line or plane that is the least-squares approximation of a set of data points is as significant as possible. As Figure 15 shows, PCA creates a visualization of data that minimizes residual variance in the least-squares sense and maximizes the variance of the projection coordinates.

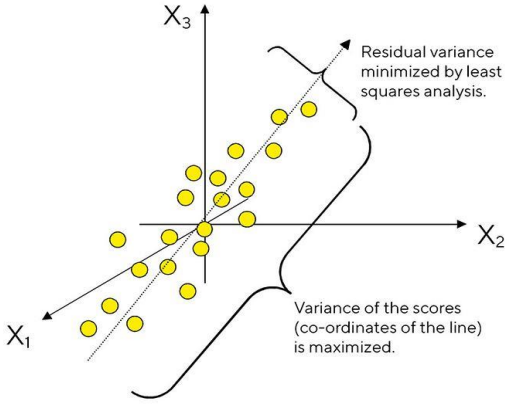

Figure 15: PCA Visualization

To sum up, the idea of PCA is to reduce the number of variables in a data set while preserving as much information as possible.

The intuition behind PCA is described as follows. Imagine a 2-dimensional dataset. A feature column represents each dimension, as Figure 16 shows.

Feature 1	Feature 2
4	2
6	3
13	6
...	...

Figure 16: A 2-Dimensional Dataset

The same dataset can be represented as a scatterplot, as Figure 17 shows.

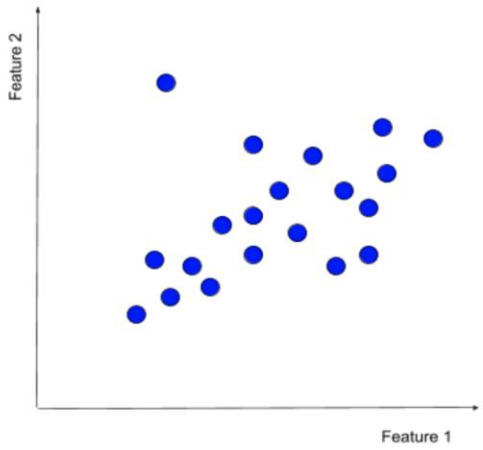

Figure 17: Scatterplot of a 2-Dimensional Dataset

PCA's main aim is to find principal components that can describe the data points with a set of "well" principal components, as Figure 18 shows.

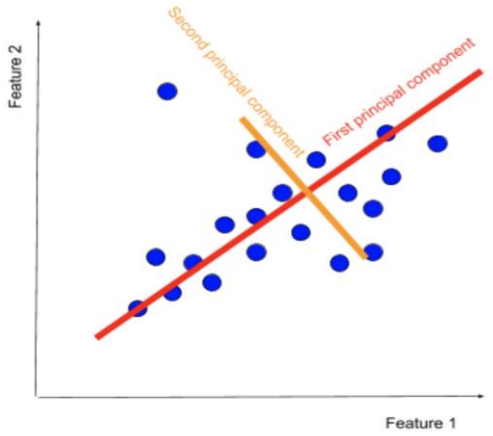

Figure 18: Representation of a 2-Dimensional Dataset with Principal Components

The principal components are vectors, but they are not picked at random. The first principal component is computed to explain the highest variance in the original features. The second component is orthogonal, explaining the highest variation left after the first principal component.

Feature vectors can be used to represent the original data. PCA lets us take things further and represent the data as linear combinations of principal components. Obtaining principle components equals performing a linear data transformation from the feature1 x feature2 axis to the PCA1 x PCA2 axis.

4.3.1.2 Importance of PCA

We do not gain much by applying PCA in the small 2-dimensional example above since a feature vector of the form (feature1, feature2) is highly similar to a vector of the form (the first principal component

(PCA1) and the second principal component (PCA2). However, in big datasets (where the number of dimensions can exceed 100), principal components remove noise by reducing a huge number of features to just a couple of principal components. Principal components are orthogonal projections of data onto a lower-dimensional space.

Theoretically, PCA generates as many principal components as features in the training dataset. In actuality, we do not keep all the principal components. Each consecutive principal component explains the variance left after its primary component, so picking just a few of the first components sufficiently approximates the original dataset without needing additional features.

As a result, a new set of features in principal components has emerged, with multiple practical applications.

4.3.1.3 Calculation of PCA

There are multiple ways to calculate PCA:

- Eigen decomposition of the covariance matrix
- Decomposition of the data matrix into singular values
- Eigenvalue approximation via power iterative computation
- NIPALS (nonlinear iterative partial least squares) is a method for computing nonlinear partial least squares.

The first method, eigen decomposition of the covariance matrix, is used to gain a deeper appreciation of PCA. There are a few steps in computing PCA:

- Feature standardization. Each feature is standardized to have a mean of 0 and a variation of 1. Features with values on different orders of magnitude preclude PCA from estimating the best principal components, as we describe later in assumptions and limitations.

- Obtain the covariance matrix computation. The covariance matrix is a square matrix with dimensions d x d, where d denotes "dimension" (or feature or column if our data is tabular). It displays the pairwise feature correlation for each feature.

- Calculate the covariance matrix's eigen decomposition. The covariance matrix's eigenvectors (unit vectors) and eigenvalues (scalars by which the eigenvector is multiplied) are calculated. If you want to brush up on your linear algebra, this is an excellent resource to refresh your knowledge of eigen decomposition.

- Sort the eigenvectors from the highest eigenvalue to the lowest; the eigenvector with the highest eigenvalue is the first principal component. More significant amounts of shared variance explained correspond to higher eigenvalues.

- Select the number of principal components. Select the top N eigenvectors (depending on their eigenvalues) to become the N principal components. The appropriate number of principal components is both subjective and problem-dependent. Usually, we look at the cumulative amount of shared variation

described by the combination of principal components and pick the number of components that still significantly explains the shared variance.

4.3.1.4 Advantages and Disadvantages of PCA

Using PCA in machine learning has various advantages:

- Easy to compute because PCA is based on linear algebra, which is computationally easy to solve by computers.
- It speeds up other machine learning algorithms because machine learning algorithms converge faster when trained on principal components instead of the original dataset.
- Counteracting the issues of high-dimensional data, because high-dimensional data causes regression-based algorithms to overfit easily by using PCA beforehand to lower the dimensions of the training dataset, we prevent the predictive algorithms from overfitting.

However, it also suffers from certain shortcomings.

- Principal components have low interpretability because they are linear combinations of the features from the original data, but they are not as easy to interpret. For instance, it is difficult to tell the essential features in the dataset after computing principal components.
- Although dimensionality reduction is advantageous, the tradeoff between information loss and dimensionality reduction comes with a cost. The loss of information is an unavoidable part of PCA. Balancing the tradeoff between dimensionality reduction and information loss is an unavoidable compromise we must make when employing PCA.

4.3.1.5 Assumptions and Limitations of PCA

Because PCA is related to the Pearson correlation's set of operations, it inherits some of the same assumptions and limitations:

- Assumes that features are correlated. PCA will not be able to determine principal components if the features (or dimensions or columns in tabular data) are not connected.
- The scale of the features has an impact on PCA. Consider two features: one that accepts values between 0 and 1000 and the other that accepts values between 0 and 1. Regardless of the actual maximum variance in the data, PCA will strongly favour the first feature as the first principal component. As a result, it is critical first to standardize the values.
- Not robust to outliers. As mentioned above, the algorithm will be biased in datasets with solid outliers. This is why removing outliers before performing PCA is advised.
- Assumes that features have a linear relationship. The algorithm does not adequately capture nonlinear relationships. As a result, it is a good idea to use standard methods like log transforms to convert nonlinear features or relationships between features to linear ones. Missing values are frequently assumed in technical implementations. When utilizing statistical software tools to compute PCA, they

frequently assume that the feature set contains no missing values (no empty rows). Remove any missing values from the rows and columns, or approximate missing values with a close approximation, for example, the mean of the column.

4.3.1.6 PCA Uses

The algorithm can be used on its own, or it can serve as a data cleaning or data preprocessing technique used before another machine learning algorithm. On its own, PCA is used across a variety of use cases:

- Visualize multidimensional data. Data visualizations are an excellent tool for communicating multidimensional data as 2- or 3-dimensional plots.
- Compress information. Principal Component Analysis compresses information to store and transmit data more efficiently. For example, it can compress images without losing too much quality or in signal processing. The technique has successfully been applied across various compression problems in pattern recognition (specifically face recognition), image recognition, and more.
- Simplify complex business decisions. PCA has been employed to simplify traditionally complex business decisions. For example, traders use over 300 financial instruments to manage portfolios. The algorithm has proven successful in the risk management of interest rate derivative portfolios, lowering the number of financial instruments from more than 300 to just 3-4 principal components.
- Clarify convoluted scientific processes. The algorithm has been applied extensively in understanding convoluted and multidirectional factors, which increase the Probability of neural ensembles triggering.

When PCA is used as part of preprocessing, the algorithm is applied to:

- Reduce the number of dimensions in the training dataset.
- De-noise the data. Because PCA is computed by finding the components which explain the most significant amount of variance, it captures the signal in the data and omits the noise.

4.3.2 Clustering Analysis

4.3.2.1 What Cluster Analysis is

Cluster analysis is a statistical method used to categorize related things into respective categories. The purpose of performing a cluster analysis is to sort different objects or data points into groups so that the degree of relationship between two objects if they belong to the same group is high and if they belong to distinct groupings is low.

Cluster analysis differs from other statistical methods since it is mainly used when researchers do not have an assumed principle or fact they are using as the foundation of their research. This analysis technique is often conducted during the exploratory phase of research since, unlike techniques such as factor analysis, it does not distinguish between dependent and independent variables. Instead, cluster

analysis is leveraged mainly to uncover structures in data without providing an explanation or interpretation.

In other words, cluster analysis uncovers data structures without elucidating their existence. For example, specific groupings within a population can be identified when cluster analysis is used in market research. These groups can then be analyzed to see how likely a population cluster is to buy products or services. A marketing team can target varying clusters with customized, targeted communication if these groupings are clearly defined.

4.3.2.2 Common Applications of Cluster Analysis

Below are the typical applications of cluster analysis.

Marketing: Marketers frequently utilize cluster analysis to define market groupings, allowing for better product positioning and communications. This allows a company to better position itself, explore new markets and develop products that specific clusters find relevant and valuable.

Insurance: Insurance companies commonly employ cluster analysis if there are many claims in one region. This allows them to figure out what is behind the increase in claims.

Geology: Geologists employ cluster analysis to assess seismic risk and potential weaknesses in earthquake-prone regions for cities along fault lines. Residents can do their best to prepare for any damage by studying the findings of this study.

4.3.2.3 What Clustering Process Look Like

The clustering process includes the below steps.

Step 1: Build and Distribute a Survey

Your survey should include numerous measures of propensity to purchase and the preferences for the product at hand. It should be delivered to your population of interest, and the sample size should be large enough to make scientifically accurate decisions.

Step 2: Analyze Response Data

It is regarded as best practice to perform a factor analysis on your survey to reduce the factors being clustered. If, after your factor analysis, it is found that a few of the questions are measuring the same thing, you should combine these questions before completing your cluster analysis.

After decreasing your data by factoring, perform the cluster examination, choose how many clusters appear fitting, and record those cluster assignments. You will presently be able to see the implications of all your factors over clusters.

Step 3: Take Informed Action

Comb through the data to identify differences in the means of factors, then name your clusters based on these discrepancies. These distinctions across clusters are, therefore, able to inform your marketing, allowing you to target exact groups of clients with the appropriate message, at the right time and in the proper manner.

4.3.2.4 Types of Cluster Analysis

There are three basic methods used to perform cluster analysis.

Hierarchical Cluster: This is the most commonly used clustering method. It builds a series of models with cluster solutions ranging from 1 to n (all cases in all clusters). This approach also works with variables rather than cases. Hierarchical clustering can group variables like factor analysis. Finally, hierarchical cluster analysis can deal with nominal, ordinal, and scale data. However, remember that you should not combine multiple levels of measurement in your study.

K-Means Cluster: This method is used to cluster massive datasets quickly. Here, researchers define the number of clusters before performing the actual study. This method helps compare multiple models with a different number of clusters assumed.

Two-Step Cluster: This method identifies groupings using a clustering algorithm by first pre-clustering and then performing hierarchical methods. Two-step clustering is best for managing larger datasets that would otherwise take too long a time to calculate with strictly hierarchical methods. Essentially, two-step cluster analysis combines hierarchical and k-means cluster analysis, so it can handle both scale and ordinal data and automatically selects the number of clusters.

4.3.2.5 Cluster Analysis Uses

Clustering allows researchers to identify and define patterns between data elements. Revealing these patterns between data points helps to distinguish and outline structures that might not have been apparent before but which give significant meaning to the data once they are discovered. Once a clearly defined structure emerges from the dataset, informed decision-making becomes much more accessible.

The Cluster analysis divides data into groups, commonly called clusters, depending on their similitudes. The objects inside the groups may be similar but dissimilar to objects in the other clusters. Clustering can be formulated as a multi-objective optimization problem. The distance function, density threshold, and the number of expected clusters commonly influence the arguments for the division into clusters.

If new input is entirely dissimilar from the observation in the clusters, it is considered an outlier and related to suspicious fraud behavior. This method can be used to discover structures in data without providing an explanation or interpolations. This means that Cluster Analysis discovers data structures without explaining why they exist.

4.4 Hybrid Methods

Hybrid methods, combining supervised and unsupervised methods, have been developed by several researchers. When an unsupervised method is followed by a supervised method, the objective is usually to discover knowledge hierarchically.

4.4.1 Self-Organizing Map and Neural Networks

The principles of the two methods are identical. A dimensionality reduction method is a Self-Organizing Map (SOM), a type of ANN trained using unsupervised learning to produce a low-dimensional, typically two-dimensional, referred to as a map, discretized representation of the input space of the training samples. Self-organizing maps are distinct from other artificial neural networks. They use competitive learning rather than error-correction learning, such as backpropagation with gradient descent, and use a neighbourhood function to retain the input space's topological properties.

After some experimental results, the conclusions were that both methods had good performance even working differently. The idea of this method is to combine the supervised and the unsupervised neural networks approach. The training data is initially divided into four classes indicating different possibilities of fraud. After applying the Neural Network method, SOM was employed to refine the training data. The classifications used were the ones obtained by the SOM method and those given by domain experts.

4.4.2 Clustering Analysis and Decision Tree

This method, developed to detect insurance subscribers' fraud, has a divide and conquer strategy separated into three steps. The first is constructing a raw and unsupervised clustering of insurance subscribers' profiles. The second creates a decision tree for each group and then converts it into a set of rules. The third evaluates each rule by establishing a mapping from the rule to a measure of its significance using summary statistics. This method can reduce the rules generated by the decision tree.

4.3 Chapter Summary

Chapter 4 focused on the methods of machine learning in fraud detection that are commonly used. There are classified in supervised, unsupervised and hybrid methods.

CHAPTER 5

Application

5.1 Introduction

Frauds committed by providers are one of the biggest problems facing Medicare. According to the government, the total Medicare spending increases exponentially due to fraud in Medicare claims. Healthcare fraud is an organized crime that involves peers of providers, physicians, and beneficiaries acting together to make fraud claims.

Rigorous analysis of Medicare data has yielded many physicians who indulge in fraud. They adopt ways in which an ambiguous diagnosis code is used to adopt the costliest procedures and drugs. Insurance companies are the most vulnerable institutions impacted because of these bad practices. Due to this reason, insurance companies increased their insurance premiums, and as result healthcare is becoming a costly matter day by day.

Healthcare fraud and abuse take many forms. As in the previous chapter mentioned, some of the most common types of fraud by providers are:

- Billing for services that were not provided
- Duplicate submission of a claim for the same service
- Misrepresenting the service provided
- Charging for a more complex or expensive service than was actually provided
- Billing for a covered service when the service provided was not covered.

5.2 Aim of the project

The goal of this chapter is to predict the potentially fraudulent providers based on the claims. We will use selected statistical techniques to build a model which can detect with lower cost and expenditure fraud instead of searching if a case is fraudulent or not separately. Along with this, important features that help detect the behavior of potentially fraudulent providers will be discovered. Furthermore, there is an exploratory for fraudulent patterns in the provider's claims to understand the future behavior of providers.

5.3 Dataset Overview

The dataset used in this project is “Healthcare provider fraud detection analysis” which was released in Kaggle, a community of data scientists and machine learners. It contains records of transactions made by potentially fraudulent providers that occurred in Medicare. The dataset contains 5410 transactions out of which only 506 are fraudulent. The dataset is highly imbalanced as the positive class accounts for only 9.35% of the total transactions. The imbalanced class distribution can be visualized in a bar diagram given in Figure 19.

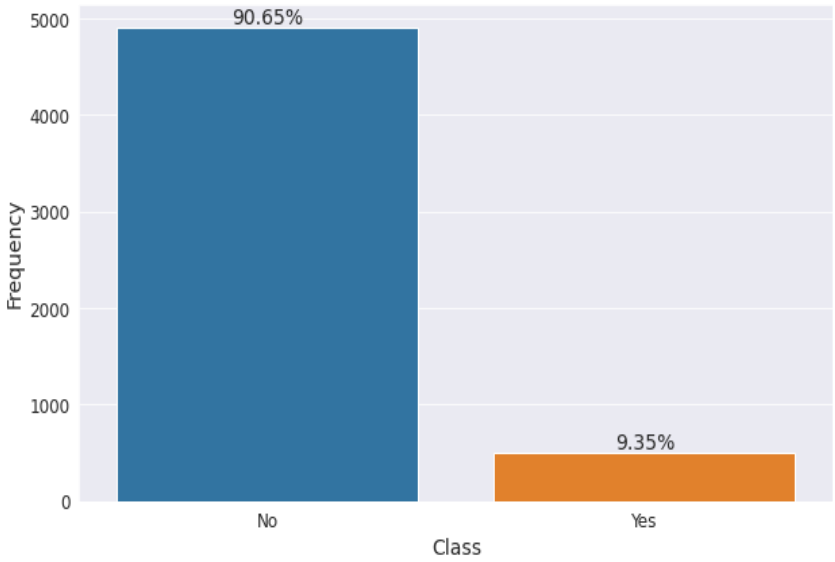

Figure 19: Visualization of Highly Imbalanced Class Distribution

The dataset includes individual datasets. The first is the Train dataset which consists of provider numbers and corresponding whether this provider is potentially fraudulent with Provider ID being the primary key in that table. The Test dataset consists of only the provider number. The Inpatient Dataset provides insights into the claims filed for those patients who are admitted to hospitals. It also provides additional details like their admission and discharge dates and admit diagnosis code. The Outpatient Dataset provides details about the claims filed for those patients who visit hospitals and are not admitted to them. It has all columns of Inpatient Data except AdmissionDt, DischargeDt, and DiagnosisGroupCode which are details recorded for patients who require hospitalization. Beneficiary Details Dataset which contains beneficiary Know Your Customer details like health conditions, DOB, race, and region they belong to. Table 2 shows in detail the descriptions of features for all datasets.

Feature	Description	Dataset Origin	Type
PotentialFraud	Indicates if the provider is potentially fraudulent	Train	Integer
Provider	It consists the ID of provider	Inpatient, Outpatient, Train, Test	String
AdmissionDt	It contains the date on which the patient was admitted into the hospital in yyyy-mm-dd format	Inpatient	Date
DischargeDt	It contains the date on which the patient was discharged from the hospital in yyyy-mm-dd format	Inpatient	Date
ClaimEndDt	It contains the date when the claim ended in yyyy-mm-dd format	Inpatient, Outpatient	Date
ClaimStartDt	It contains the date when the claim started in yyyy-mm-dd format	Inpatient, Outpatient	Date
AttendingPhysician	It contains the ID of the Physician who attended the patient	Inpatient, Outpatient	String
OperatingPhysician	It contains the ID of the Physician who operated on the patient.	Inpatient, Outpatient	String
OtherPhysician	It contains the ID of the Physician other who treated the patient	Inpatient, Outpatient	String
ClmAdmitDiagnosisCode	Diagnosis code indicating the beneficiary's initial diagnosis at admission	Inpatient, Outpatient	String
ClmDiagnosisCode_1	Indicates if the patient has existed that diagnosis code	Inpatient, Outpatient	String
ClmDiagnosisCode_2	Indicates if the patient has existed that diagnosis code	Inpatient, Outpatient	String
ClmDiagnosisCode_3	Indicates if the patient has existed that diagnosis code	Inpatient, Outpatient	String
ClmDiagnosisCode_4	Indicates if the patient has existed that diagnosis code	Inpatient, Outpatient	String
ClmDiagnosisCode_5	Indicates if the patient has existed that diagnosis code	Inpatient, Outpatient	String
ClmDiagnosisCode_6	Indicates if the patient has existed that diagnosis code	Inpatient, Outpatient	String
ClmDiagnosisCode_7	Indicates if the patient has existed that diagnosis code	Inpatient, Outpatient	String
ClmDiagnosisCode_8	Indicates if the patient has existed that diagnosis code	Inpatient, Outpatient	String
ClmDiagnosisCode_9	Indicates if the patient has existed that diagnosis code	Inpatient, Outpatient	String
ClmDiagnosisCode_10	Indicates if the patient has existed that diagnosis code	Inpatient, Outpatient	String
DiagnosisGroupCode	It contains a group code for the diagnosis done on the patient	Inpatient, Outpatient	String
ClmProcedureCode_1	Indicates if the patient has existed that procedure code	Inpatient, Outpatient	String
ClmProcedureCode_2	Indicates if the patient has existed that procedure code	Inpatient, Outpatient	String
ClmProcedureCode_3	Indicates if the patient has existed that procedure code	Inpatient, Outpatient	String
ClmProcedureCode_4	Indicates if the patient has existed that procedure code	Inpatient, Outpatient	String
ClmProcedureCode_5	Indicates if the patient has existed that procedure code	Inpatient, Outpatient	String
ClmProcedureCode_6	Indicates if the patient has existed that procedure code	Inpatient, Outpatient	String
DeductibleAmtPaid	It consists of the amount by the patient. That is equal to the total claim amount minus the reimbursed amount	Inpatient, Outpatient	Float
InscClaimAmtReimbursed	It contains the amount reimbursed for that particular claim	Inpatient, Outpatient	Float
BeneID	It contains the unique id of each beneficiary	Inpatient, Outpatient, Beneficiaries	String
ClaimID	It contains the unique id of the claim submitted by the provider	Inpatient, Outpatient, Beneficiaries	String
County	It contains the country of the beneficiary	Beneficiaries	Integer
Race	It contains the race of the beneficiary	Beneficiaries	Integer
State	It contains the state of the beneficiary	Beneficiaries	Integer
Gender	It contains the gender of the beneficiary	Beneficiaries	Integer
DOB	Date of birth of beneficiary	Beneficiaries	Date
DOD	Date of death of beneficiary	Beneficiaries	Date
RenalDiseaseIndicator	It contains if the patient has existing kidney disease	Beneficiaries	Integer
ChronicCond_Heartfailure	Indicates if the patient has existed that disease, which also indicates the risk score of that patient	Beneficiaries	Integer
ChronicCond_stroke	Indicates if the patient has existed that disease, which also indicates the risk score of that patient	Beneficiaries	Integer
ChronicCond_Alzheimer	Indicates if the patient has existed that disease, which also indicates the risk score of that patient	Beneficiaries	Integer
ChronicCond_Cancer	Indicates if the patient has existed that disease, which also indicates the risk score of that patient	Beneficiaries	Integer
ChronicCond_Diabetes	Indicates if the patient has existed that disease, which also indicates the risk score of that patient	Beneficiaries	Integer
ChronicCond_IschemicHeart	Indicates if the patient has existed that disease, which also indicates the risk score of that patient	Beneficiaries	Integer
ChronicCond_KidneyDisease	Indicates if the patient has existed that disease, which also indicates the risk score of that patient	Beneficiaries	Integer
ChronicCond_ObstrPulmonary	Indicates if the patient has existed that disease, which also indicates the risk score of that patient	Beneficiaries	Integer
ChronicCond_rheumatoidarthritis	Indicates if the patient has existed that disease, which also indicates the risk score of that patient	Beneficiaries	Integer
ChronicCond_Depression	Indicates if the patient has existed that disease, which also indicates the risk score of that patient	Beneficiaries	Integer
ChronicCond_Osteoporasis	Indicates if the patient has existed that disease, which also indicates the risk score of that patient	Beneficiaries	Integer
NoOfMonths_PartACov	Medicare has four parts. Part A provides coverage for hospital care	Beneficiaries	Integer
NoOfMonths_PartBCov	Medicare has four parts. Part B provides outpatient medical services (ie, doctor visits)	Beneficiaries	Integer
IPAnnualDeductibleAmt	It consists of a premium paid by the patient for hospitalization annually	Beneficiaries	Float
IPAnnualReimbursementAmt	It consists of the maximum reimbursement amount for hospitalization annually	Beneficiaries	Float
OPAnnualDeductibleAmt	It consists of a premium paid by the patient for outpatient visits annually	Beneficiaries	Float
OPAnnualReimbursementAmt	It consists of the maximum reimbursement amount for outpatient visits annually	Beneficiaries	Float

Table 2: Description of Datasets

5.4 Data Pre-processing

5.4.1 Merge Datasets

As already mentioned, the project aims to predict whether a provider commits fraud or not based on the claims filed by them. Grouping and aggregating numeric features to the provider level helped in detecting the behavior of their transactions overall. Since the data of providers, beneficiaries, and claims are all segregated, it becomes necessary to combine all datasets into one. The steps we took to achieve the merging of all datasets are listed as follows.

- Merge Inpatient and Outpatient data based on common columns.
- Merge Beneficiary data with Inpatient and Outpatient data on BeneID

- Merge provider details with previously merged data on ProviderID.
- Grouping data to count BeneID and ClaimID for each ProviderID.
- Calculating the mean of values for some variables for each provider
- Calculating the sum of values for some variables for each provider
- Merge count, mean and sum datasets into one to create the final data frame.

5.4.2 Handling Types and Missing Values

Before starting with the creation of new variables the types of some variables need to be converted as also to fill the null values. The steps we took to accomplish the transformation of variables are listed as follows.

- Replacing Yes with 1 and No with 0 in the PotentialFraud column
- Changing DOB and DOD to date type
- Changing State and Race to categorical type
- One hot encoding for State and Race.
- Replacing 1 with Female and 2 with Male
- Replacing '0' with 0 and 'Y' with 1 in RenalDiseaseIndicator column
- Replacing 2 with 0 for all chronic conditions
- Changing data type of ClaimStartDt and ClainEndDt to datetime
- Changing data type of AdmissionDt and DischargeDt to datetime
- Replacing null values with 0 in DeductibleAmtPaid.

5.4.3 Adding New Variables

Adding features from grouping them helped in improving the metrics of prediction and fraud pattern recognition. Below are the new variables that were added to the data set along with their descriptions.

Feature	Description	Dataset Origin	Type
BeneIDcount	This indicates the count number of beneficiaries for each provider, which replace the variable BeneID.	Inpatient, Outpatient, Beneficiaries	Integer
ClaimIDcount	This indicates the count number of claims for each provider, which replace the variable ClaimID.	Inpatient, Outpatient	Integer
IsInpatient	This is a flag column which indicates whether is Inpatient or Outpatient with 1 meaning Inpatient and 0 meaning Outpatient.	Inpatient, Outpatient	Integer
IsClaimCode	This is a flag column which indicates the presence of ClmAdmitDiagnosisCode in a claim with 1 meaning Yes and 0 meaning No	Inpatient, Outpatient	Integer
IsDeductible	This column indicates if any deductible was paid by the beneficiary	Inpatient, Outpatient	Integer
IsGroupCode	This is a flag column which indicates the presence of DiagnosisGroupCode in a claim with 1 meaning Yes and 0 meaning No	Inpatient, Outpatient	Integer
DaysAdmitted	This indicates the number of days a beneficiary was admitted in the hospital and is calculated by subtracting admission date from discharge date	Inpatient, Outpatient	Integer
DaysClaim	This indicates the claim period of a beneficiary which is the difference between ClaimStartDt and ClaimEndDt	Inpatient, Outpatient	Integer
ExtraDaysClaim	This indicates the number of extra days a beneficiary was admitted in the hospital	Inpatient, Outpatient	Integer
TotalClaimCodes	This indicates the total number of claim diagnosis codes used in the medical bill made by the provider	Inpatient, Outpatient	Integer
TotalProcedureCodes	This indicates the total number of medical procedures used on a beneficiary which gives an estimated cost incurred by the beneficiary	Inpatient, Outpatient	Integer
TotalPhysicians	This indicates the total number of physicians who treated, attended or operating a beneficiary	Inpatient, Outpatient	Integer
DiseasesCount	This indicates the total number of chronic diseases a beneficiary who made the claim had	Beneficiaries	Integer
Age	Age of beneficiary	Beneficiaries	Integer
IsDead	This is a flag column to indicate if the beneficiary was dead when the claim was made	Beneficiaries	Integer

5.4.4 Statistical Hypothesis Tests and Correlation with target value

Table 3 shows the original and new features' correlation with the target variable along with their statistical significance.

Feature	Correlation with target feature	Statistical Significant
PotentialFraud	1.000000	Yes
Provider	-	Yes
AdmissionDt	-	Yes
DischargeDt	-	Yes
ClaimEndDt	-	Yes
ClaimStartDt	-	Yes
AttendingPhysician	-	Yes
OperatingPhysician	-	Yes
OtherPhysician	-	Yes
ClmAdmitDiagnosisCode	-	Yes
ClmDiagnosisCode_1	-	Yes
ClmDiagnosisCode_2	-	Yes
ClmDiagnosisCode_3	-	Yes
ClmDiagnosisCode_4	-	Yes
ClmDiagnosisCode_5	-	Yes
ClmDiagnosisCode_6	-	Yes
ClmDiagnosisCode_7	-	Yes
ClmDiagnosisCode_8	-	Yes
ClmDiagnosisCode_9	-	Yes
ClmDiagnosisCode_10	-	Yes
DiagnosisGroupCode	-	Yes
ClmProcedureCode_1	0.076418	Yes
ClmProcedureCode_2	0.038117	Yes
ClmProcedureCode_3	0.015423	Yes
ClmProcedureCode_4	0.006820	Yes
ClmProcedureCode_5	0.002819	No
ClmProcedureCode_6	-	No
DeductibleAmtPaid	0.112016	No
InscClaimAmtReimbursed	0.080613	No
BeneID	0.000000	Yes
BeneIDcount	0.393531	Yes
ClaimID	-	No
ClaimIDcount	0.374197	No
County	0.011551	Yes
Race	0.024486	Yes
Gender	-0.000460	No
State	-0.041872	Yes
Age	0.008210	No
DOB	-	Yes
DOD	-	Yes
RenalDiseaseIndicator	0.391002	Yes
ChronicCond_Heartfailure	0.384131	Yes
ChronicCond_stroke	0.399206	Yes
ChronicCond_Alzheimer	0.380344	Yes
ChronicCond_Cancer	0.376945	Yes
ChronicCond_Diabetes	0.378881	Yes
ChronicCond_IschemicHeart	0.380093	Yes
ChronicCond_KidneyDisease	0.394239	Yes
ChronicCond_ObstrPulmonary	0.396191	Yes
ChronicCond_rheumatoidarthritis	0.380161	Yes
ChronicCond_Depression	0.377411	No
ChronicCond_Osteoporasis	0.001181	No
NoOfMonths_PartACov	0.005022	No
NoOfMonths_PartBCov	0.001959	No
IPAnnualDeductibleAmt	0.036514	No
IPAnnualReimbursementAmt	0.035027	No
OPAnnualDeductibleAmt	0.002919	No
OPAnnualReimbursementAmt	0.002077	No
IsInpatient	0.113401	Yes
IsClaimCode	0.053984	Yes
IsDeductible	0.088046	Yes
IsGroupCode	0.113401	Yes
IsDead	-0.001325	No
DaysAdmitted	0.082004	No
DaysClaim	0.028640	No
ExtraDaysClaim	-0.009239	No
TotalClaimCodes	0.189909	No
TotalPhysicians	0.037047	No
TotalProcedureCodes	0.188194	No
DiseasesCount	0.014685	No

Table 3: Correlation and Statistical Significance of Features

5.5 Exploratory Data Analysis

5.5.1 Original Data

5.5.1.1 Distribution of Class Data

As before highlighted, this is a highly imbalanced dataset. Figure 20 shows that there are 9.35% fraudulent providers and 90.65% non-fraudulent providers.

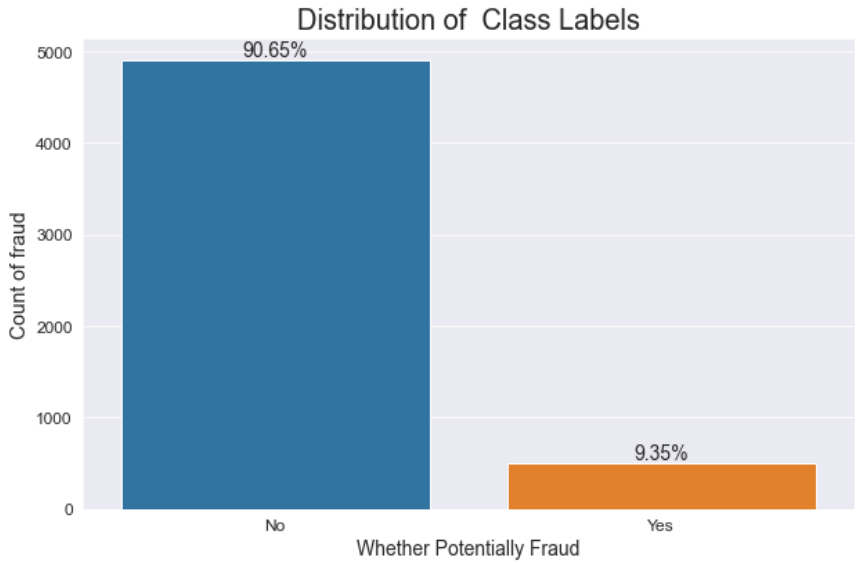

Figure 20: Distribution of Class Labels

5.5.1.2 Most Common Procedure and Diagnosis Codes of Inpatient and Outpatient

In Figure 21 we observe that procedure code 4019 is the most common procedure that is followed, around 6.6% of patients have undergone procedure code 4019, and procedures code 4019, 9904, 2724, 8154, and 66 are the top 5 procedure codes for inpatient data.

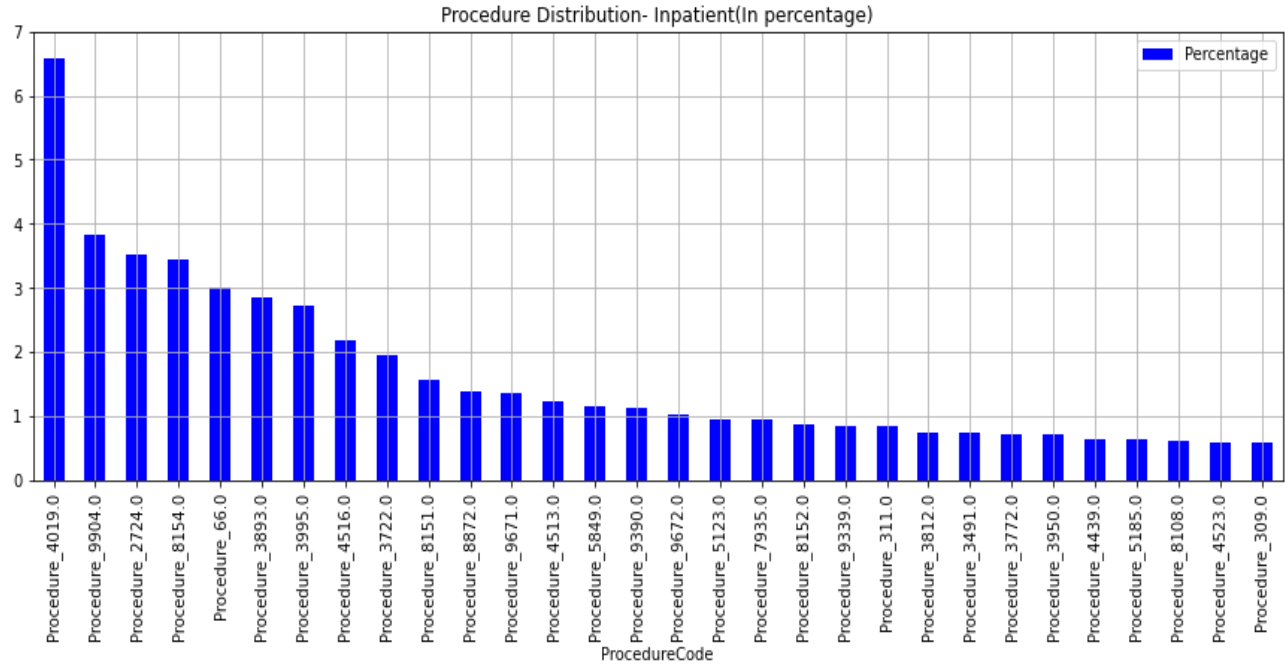

Figure 21: Most Common Procedure Codes of Inpatient

In Figure 22 we observe that diagnosis code 4019 is the most common diagnosis a patient undergoes, around 4.5% of patients have undergone Diagnosis code 4019, diagnosis codes 4019, 2724, 25000, 41401, and 4280 are the top 5 Diagnosis codes for inpatient data.

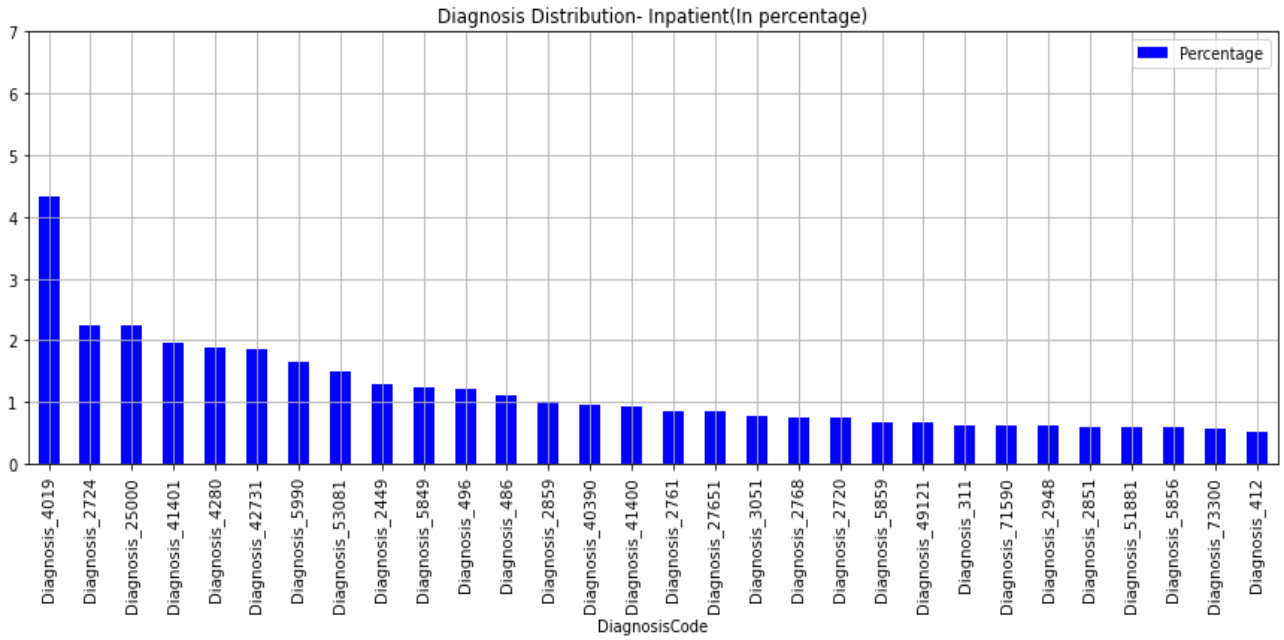

Figure 22: Most Common Diagnosis Codes of Inpatient

In Figure 23 we observe that procedure code 9904 is the most common procedure that is followed, around 7.5% of patients have undergone procedure code 9904, and procedure codes 9904, 3722, 4516, 2724, and 66 are the top 5 procedure codes for outpatient data.

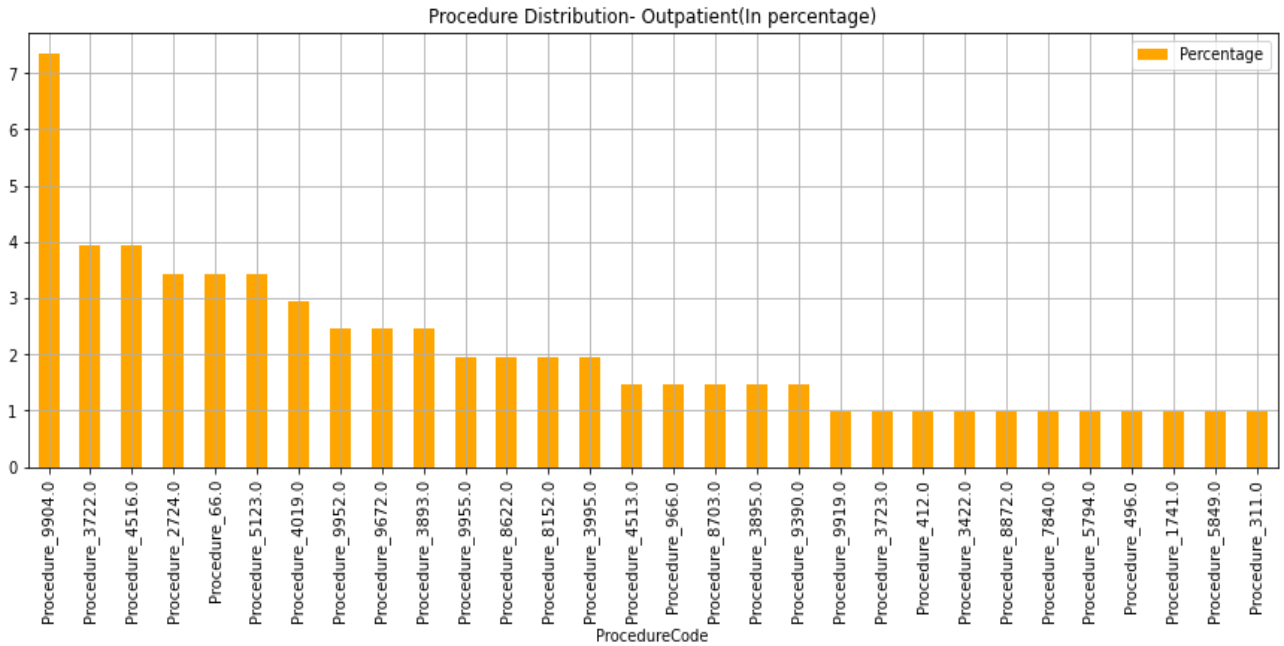

Figure 23: Most Common Procedure Codes of Outpatient

In Figure 24 we observe that diagnosis code 4019 is the most common Diagnosis a patient undergoes, around 4.8% of patients have undergone Diagnosis code 4019, and diagnosis codes 4019, 25000, 2724, V5869, and 401 are the top 5 Diagnosis codes for inpatient data.

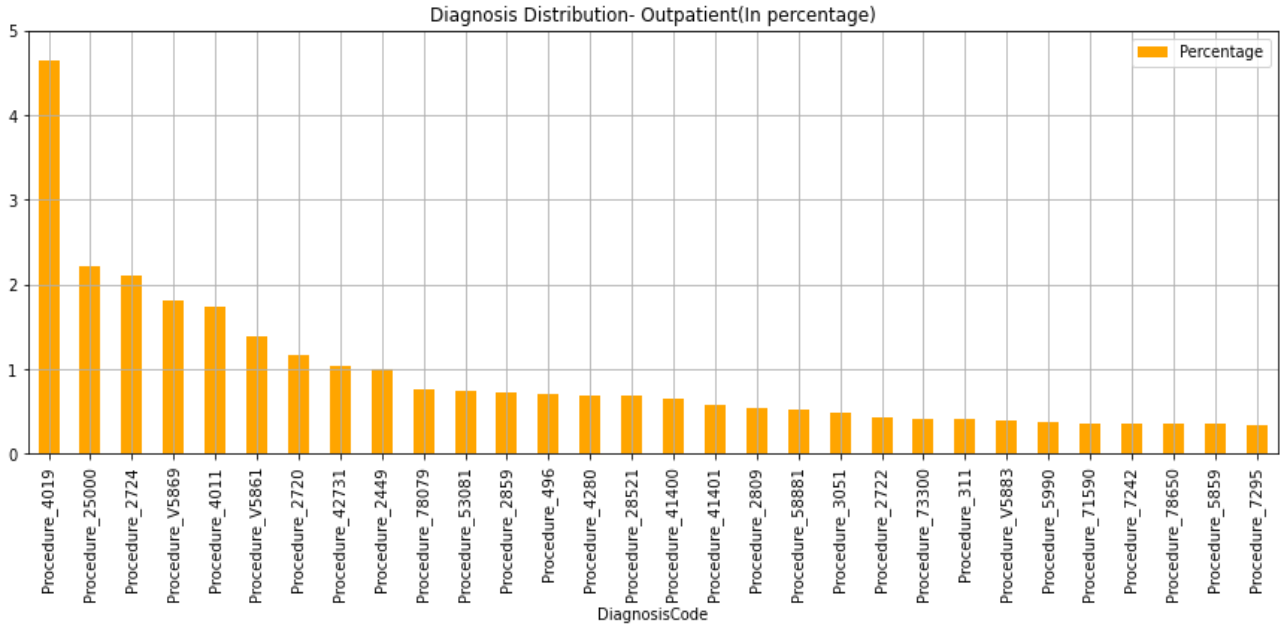

Figure 24: Most Common Diagnosis Codes of Outpatient

5.5.1.3 Most Common Codes of Physicians for Inpatient and Outpatient

In Figure 25 we observe that most inpatients are attended by physician PHY422134, and around 1% of the inpatients are attended by physician PHY422134.

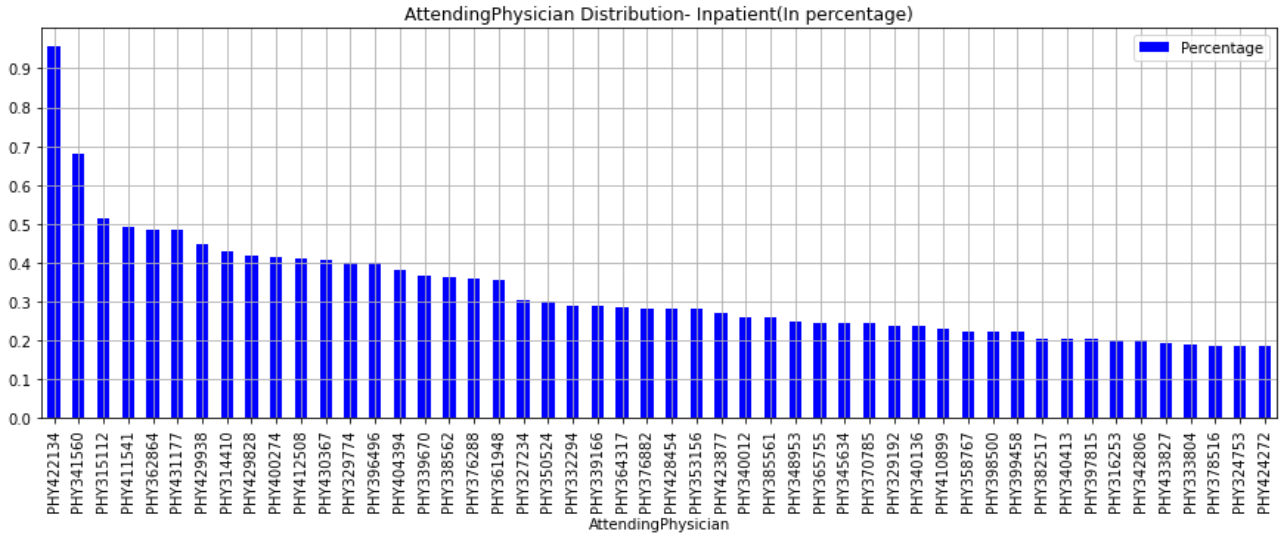

Figure 25: Most Common Codes of Attending Physicians for Inpatient

In Figure 26 we observe that physician PHY429430 performs most of the operations and around 1% of the inpatients are attended by physician PHY429430.

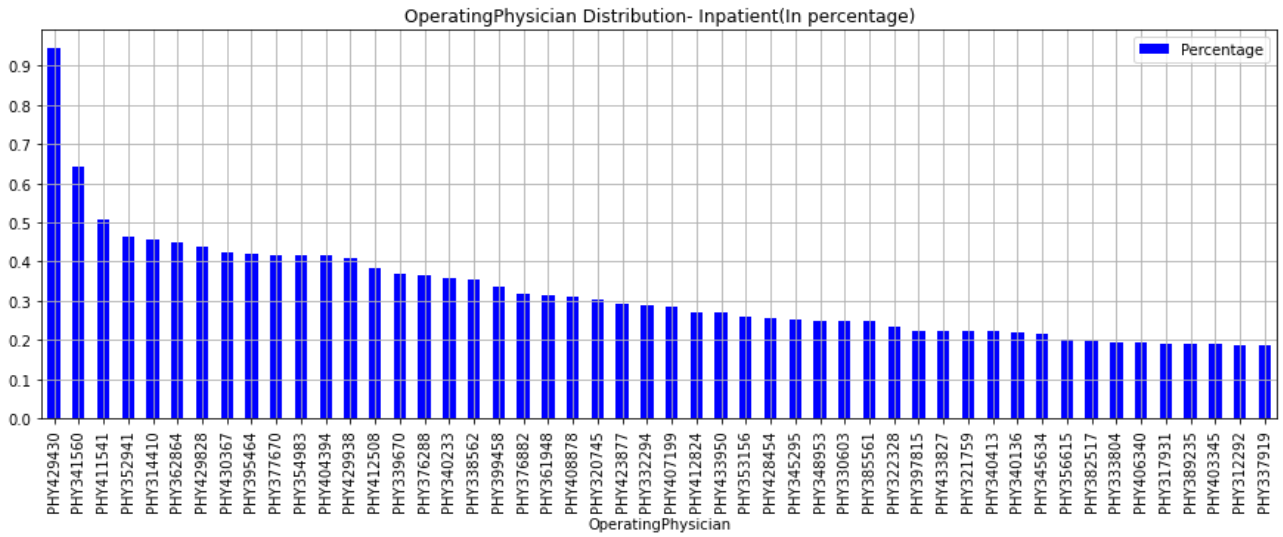

Figure 26: Most Common Codes of Operating Physicians for Inpatient

In Figure 27 we observe that the most common code of other physicians is PHY416093.

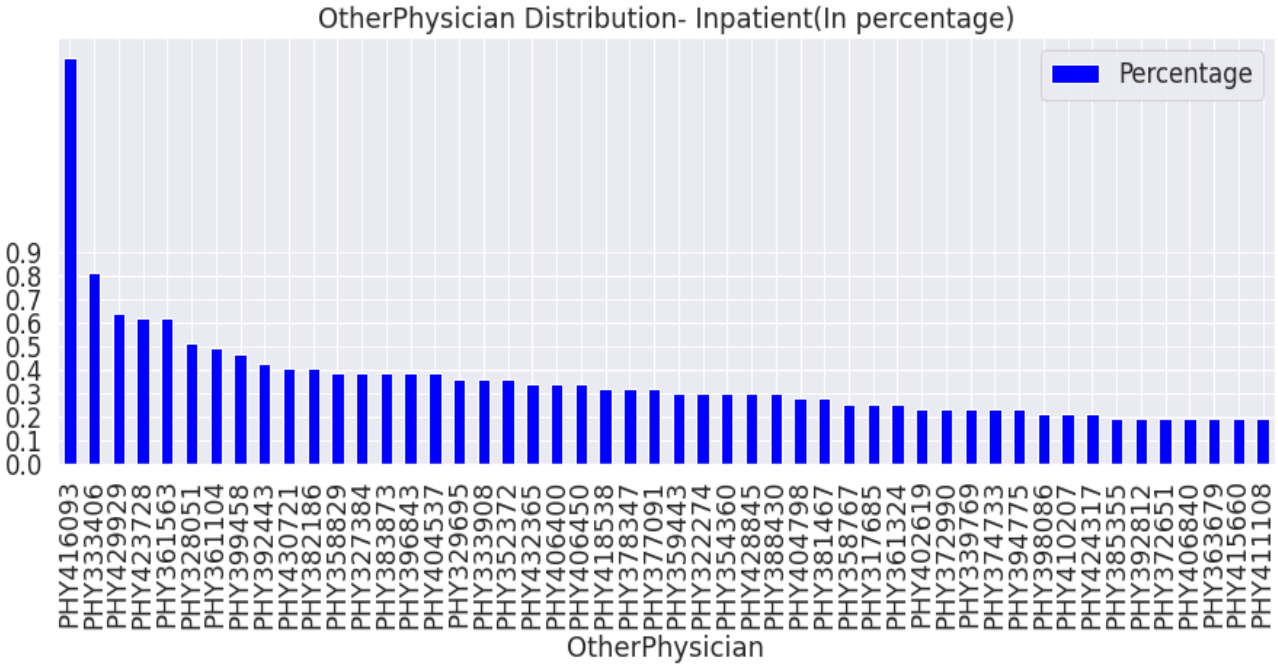

Figure 27: Most Common Codes of Other Physicians for Inpatient

In Figure 28 we observe that most patients are attended by physician PHY330576 and around 0.48% of the outpatients are attended by physician PHY330576.

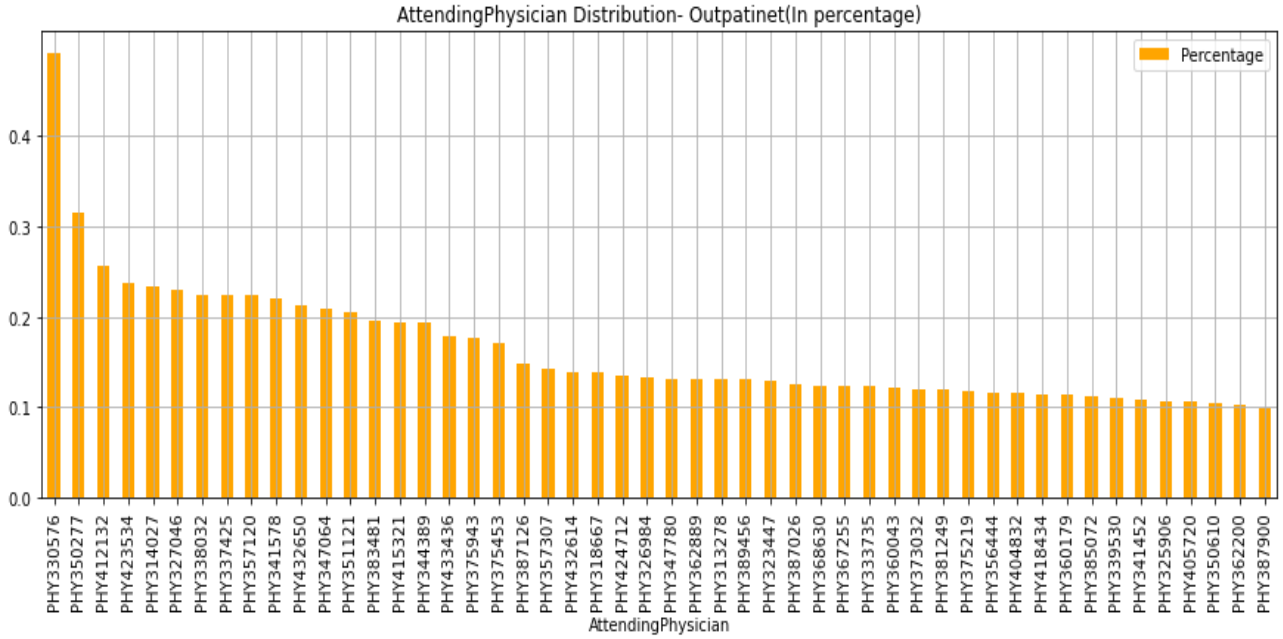

Figure 28: Most Common Codes of Attending Physicians for Outpatient

In Figure 29 we observe that physician PHY330576 performs most of the operations and around 0.48% of the outpatients are attended by physician PHY33057.

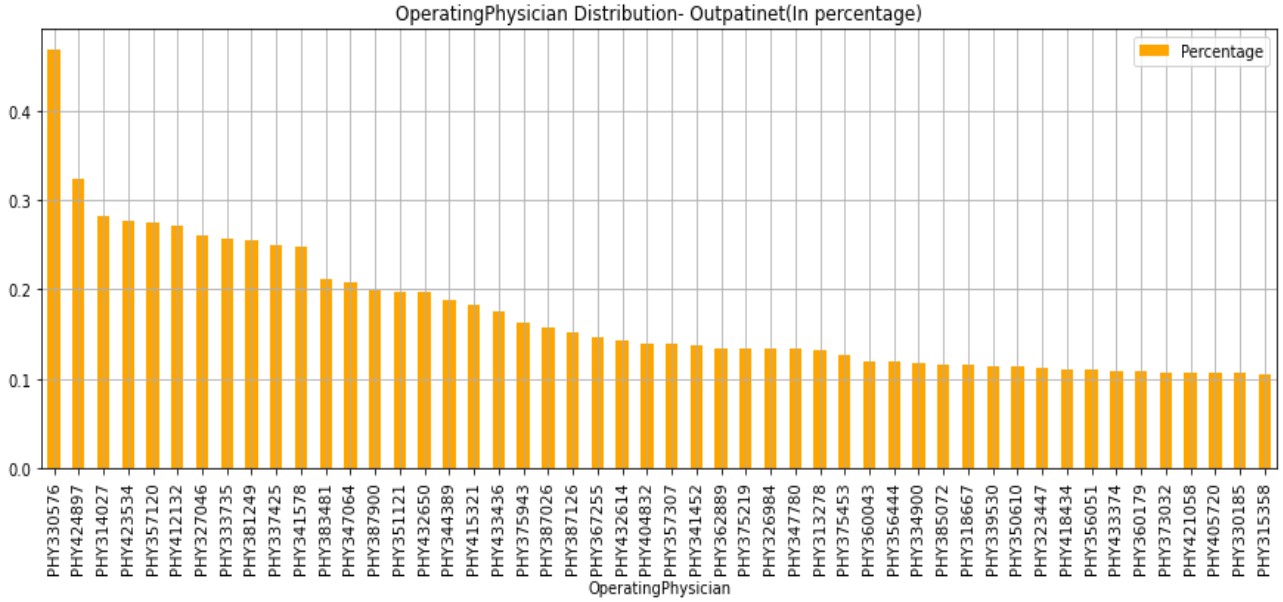

Figure 29: Most Common Codes of Operating Physicians for Outpatient

In Figure 30 we observe that the most common code of other physicians is PHY412132.

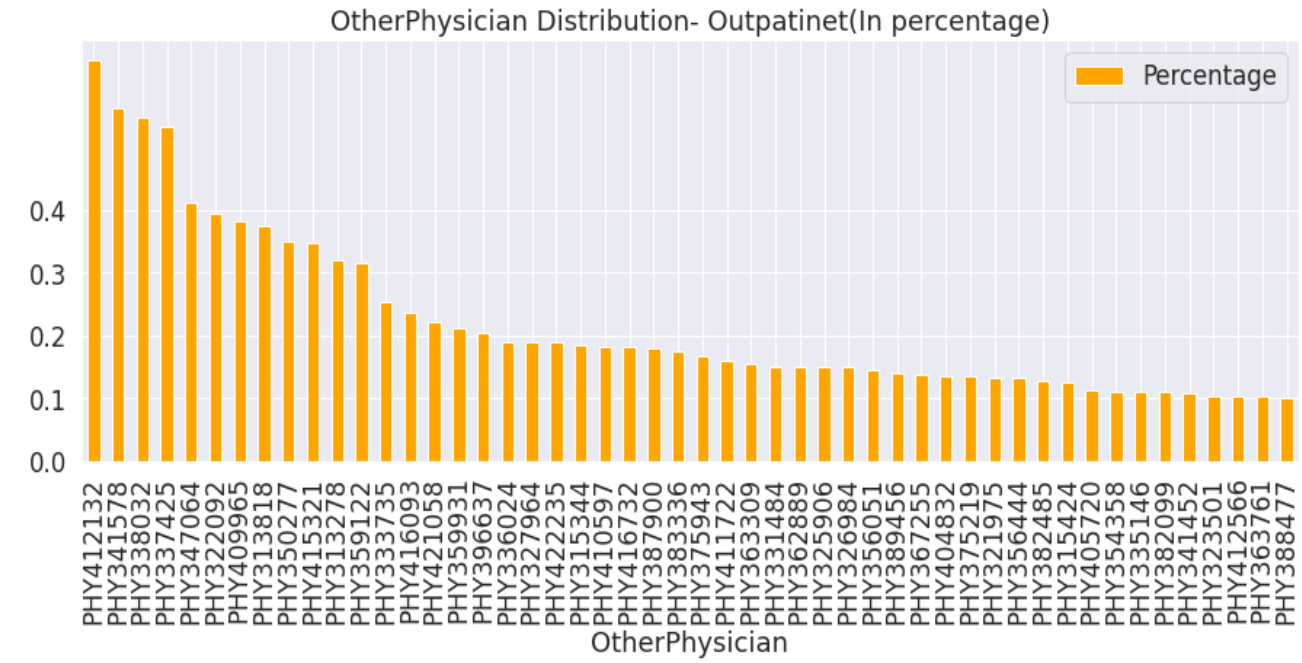

Figure 30: Most Common Diagnosis of Other Physicians for Outpatient

5.5.1.4 Most Common States, Countries, and Races of Inpatient and Outpatient

In Figures 31, 32, and 33 we observe that states coded as 5, 10, 45 and 33 are the most common an inpatient outcomes from for merge data. County coded as 10, 60, 200, and 0 are the most common an inpatient outcomes from for merge data. Race 1 is the most common an inpatient outcomes from for merge data with 84.32%.

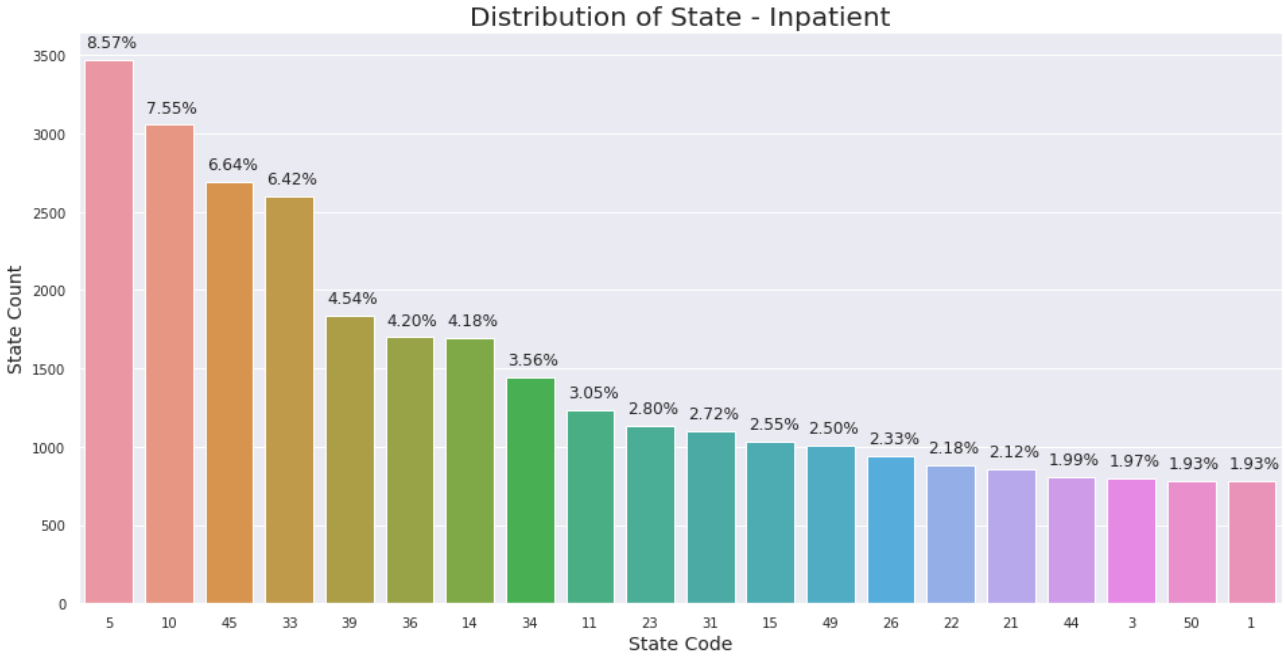

Figure 31: Most Common States of Inpatient

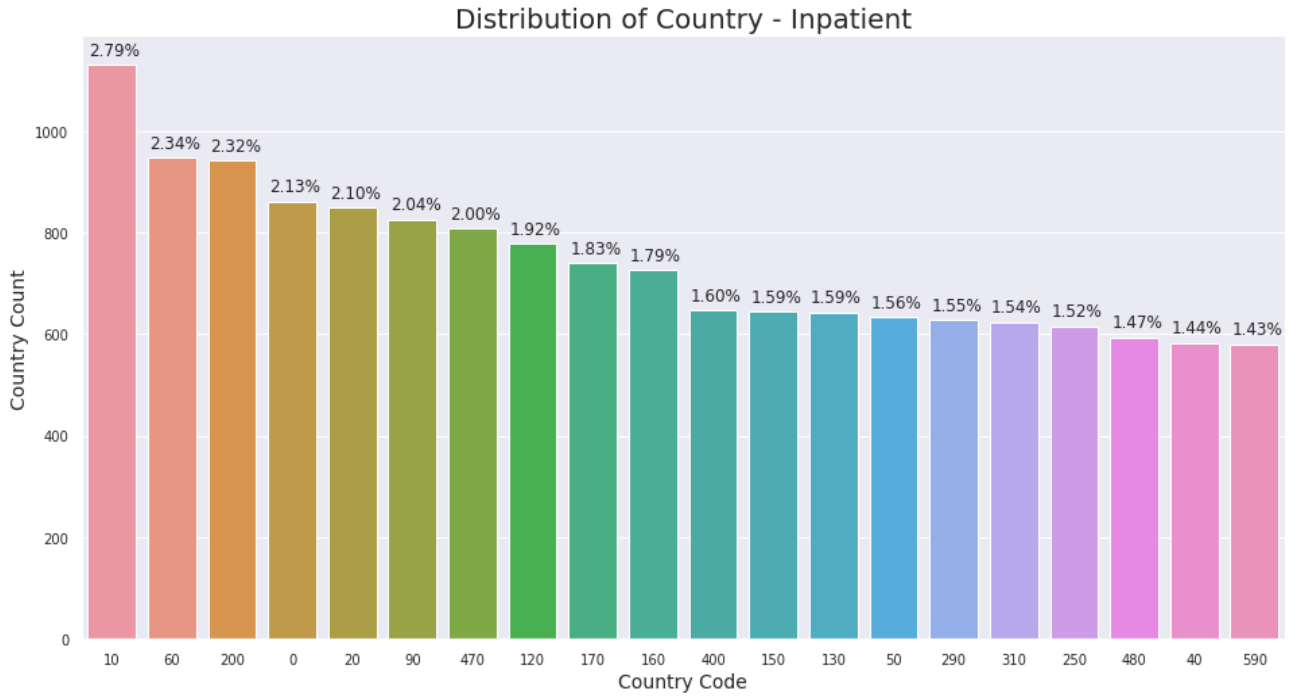

Figure 32: Most Common Countries of Inpatient

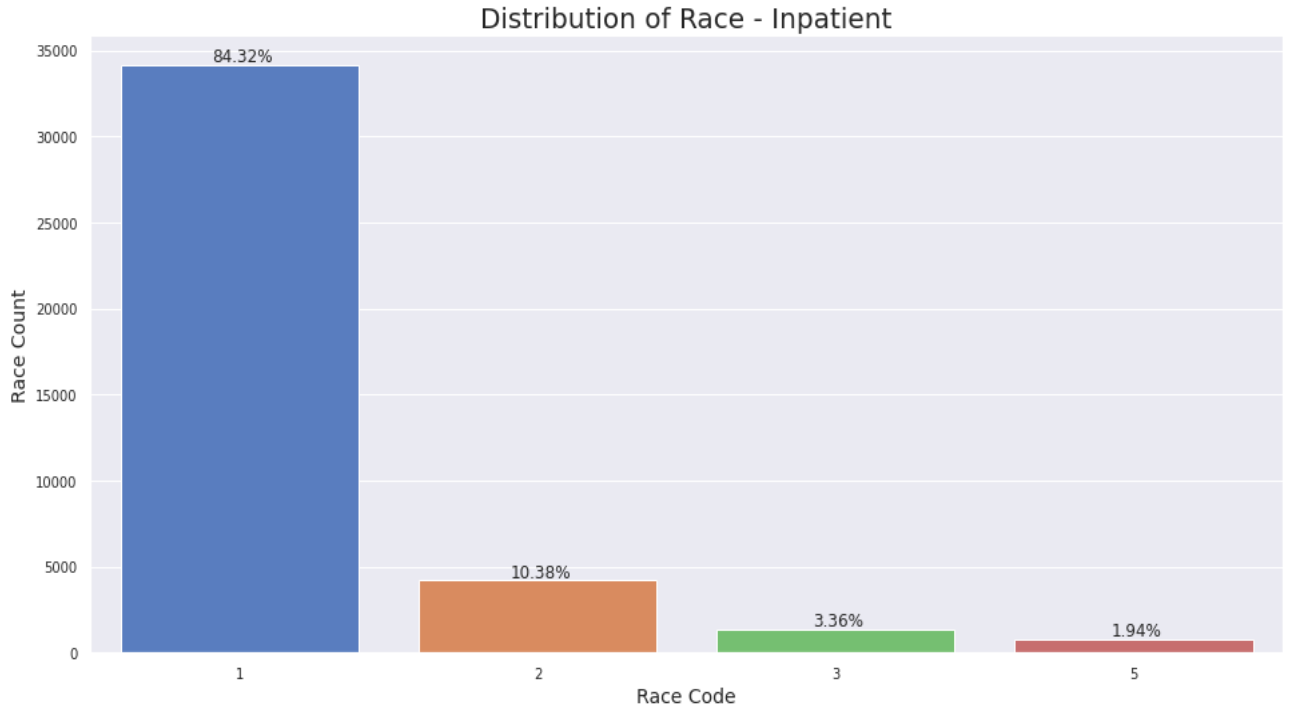

Figure 33: Most Common Races of Inpatient

In Figures 34, 35, and 36 we observe that states coded as 5, 10, 33 and 45 are the most common an outpatient outcomes from for merge data. County coded as 200, 10, 20, and 470 are the most common an outpatient outcomes from for merge data. Race 1 is the most common an outpatient outcomes from for merge data with 84.39%.

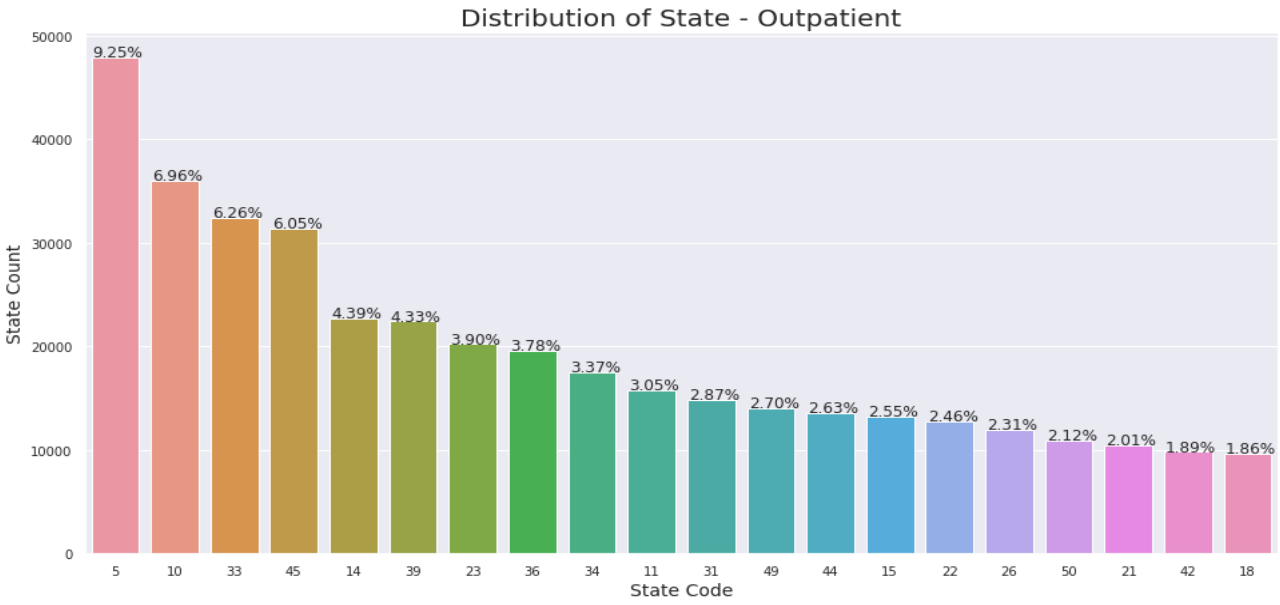

Figure 34: Most Common States of Outpatient

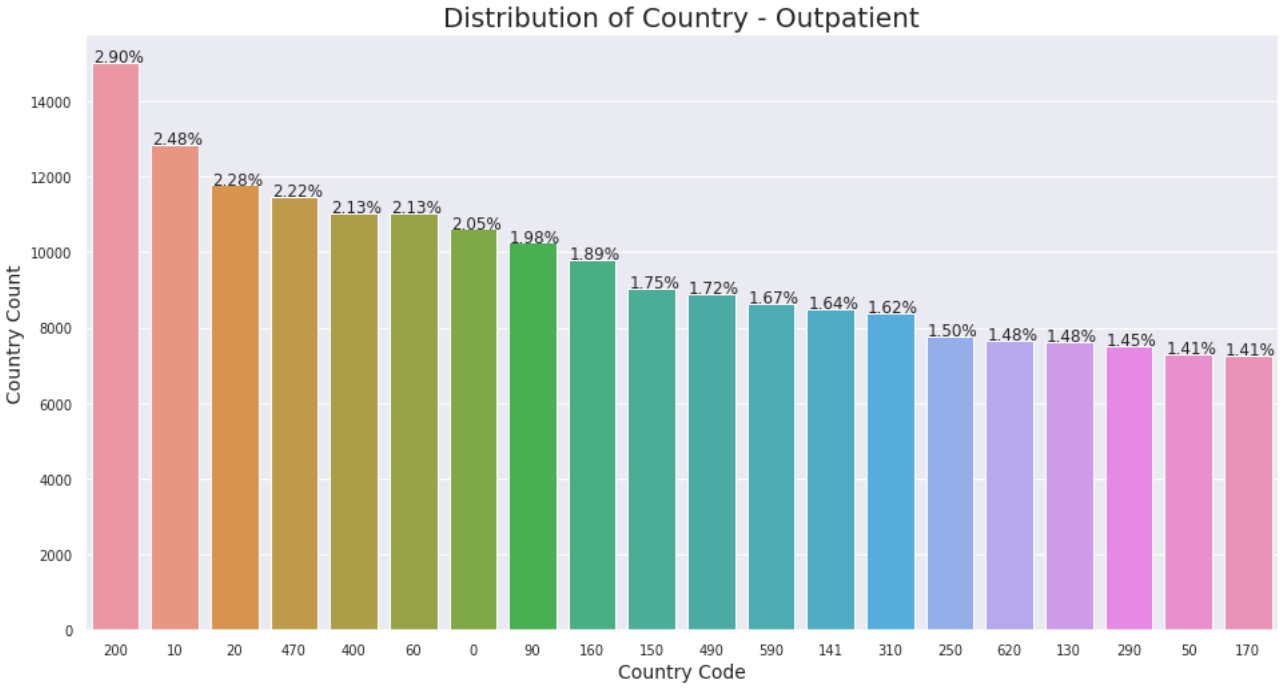

Figure 35: Most Common Countries of Outpatient

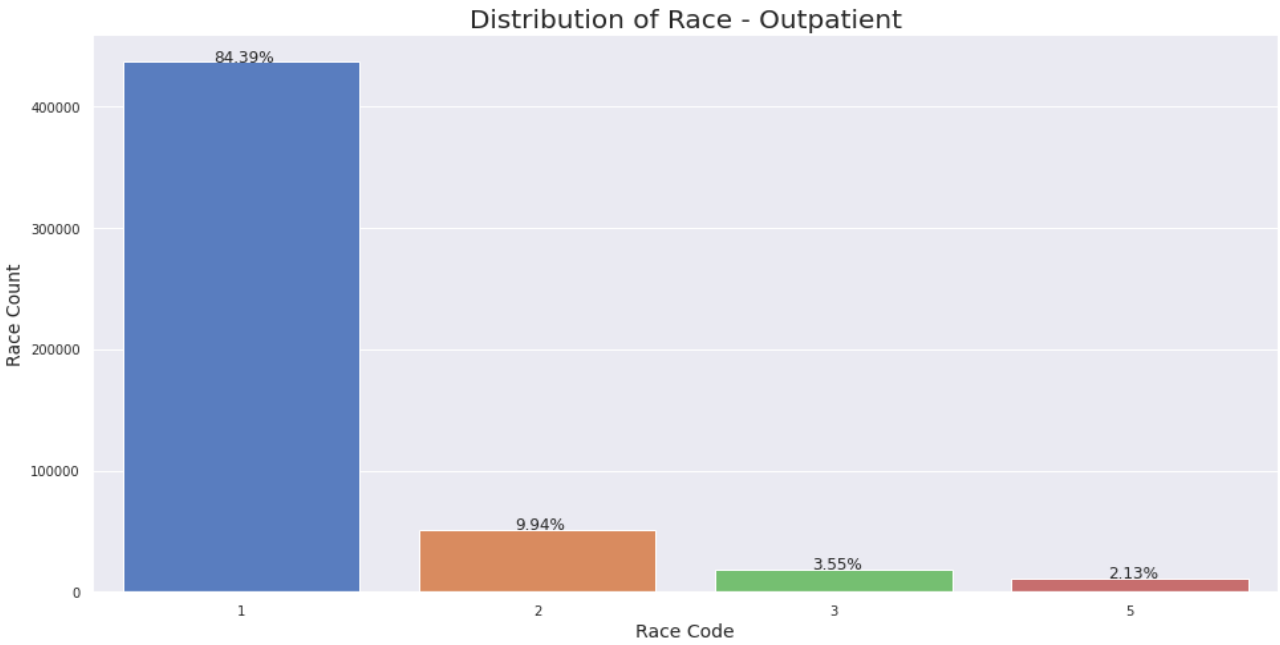

Figure 36: Most Common Races of Outpatient

5.5.1.5 Amount of Reimbursement of Inpatient and Outpatient

In Figure 37 we observe that the distribution of the amount that is paid as claim reimbursement seems like a log-normal distribution. Most of all reimbursed amount is between 0 and 20000 and in very few cases amount more than 20000 is paid for claim reimbursement. Total amount of reimbursement for inpatient is 408,297,020 and for outpatient is 148,246,120. So, that is 408 and 148 million for inpatient and outpatient respectively.

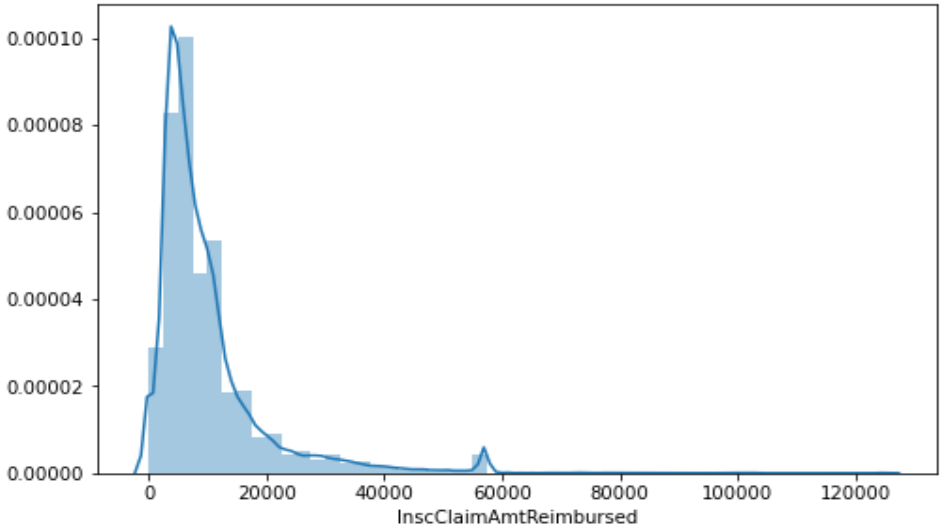

Figure 37: Reimbursement Amount of Inpatient

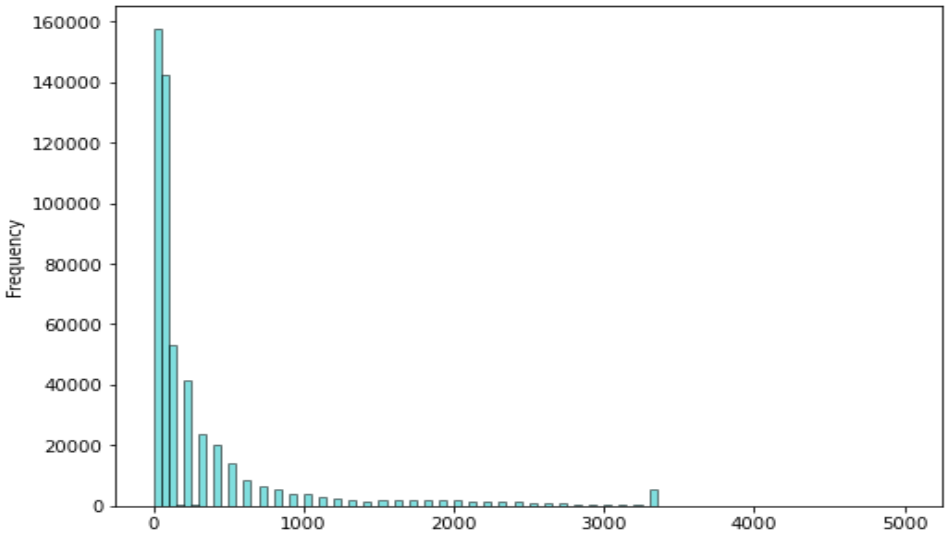

Figure 38: Reimbursement Amount of Outpatient

5.5.1.6 Fraudulent Providers

5.5.1.6.1 Percentage of Fraudulent Providers of Inpatient and Outpatient

Figure 39 shows that inpatients consist of 38,90% fraudulent providers and outpatients 61.10%.

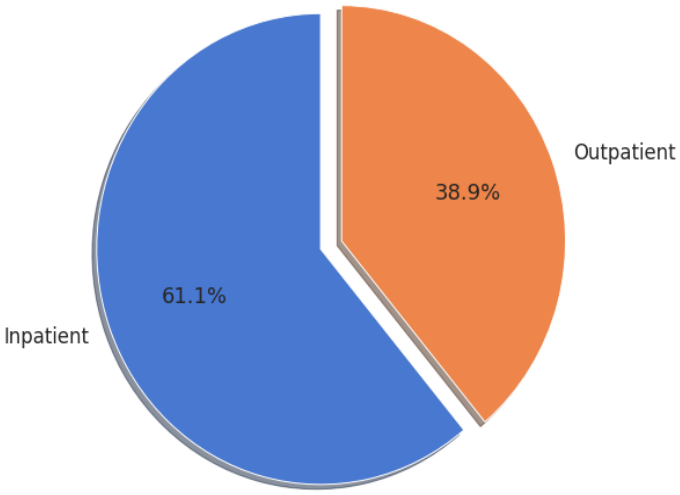

Figure 39: Percentages of Inpatient and Outpatient for Fraudulent Providers

5.5.1.6.2 Most Common Procedure and Diagnosis Codes Used by Fraudulent Providers of Inpatient and Outpatient

In Figure 40 we observe that procedure code 4019 is the most common procedure that is followed, around 6.6% of inpatients have undergone procedure code 4019, and procedure codes 4019, 2724, 9904, and 8154 are the top 5 procedure codes for inpatient data.

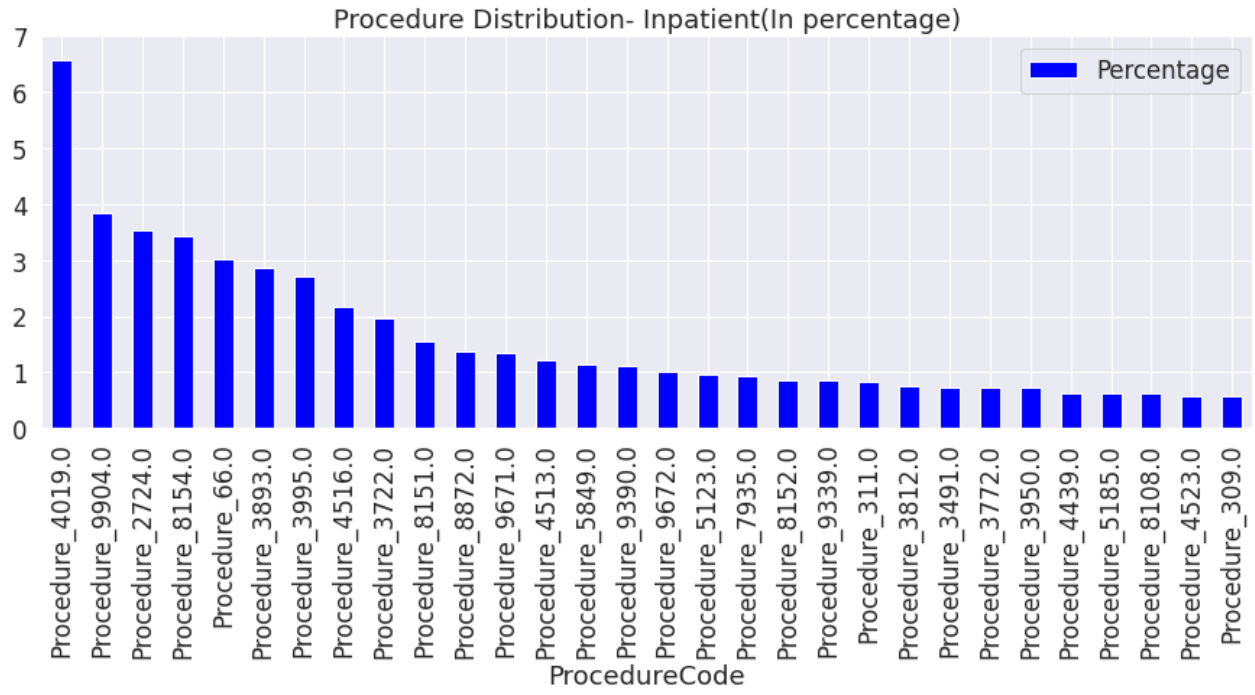

Figure 40: Most Common Procedure Codes Used by Fraudulent Providers of Inpatient

In Figure 41 we observe that diagnosis code 4019 is the most common diagnosis an inpatient undergoes, around 4.1% of inpatients have undergone diagnosis code 4019, and diagnosis codes 4019, 25000, 2724, 41401, and 4280 are the top 5 Diagnosis codes for inpatient data.

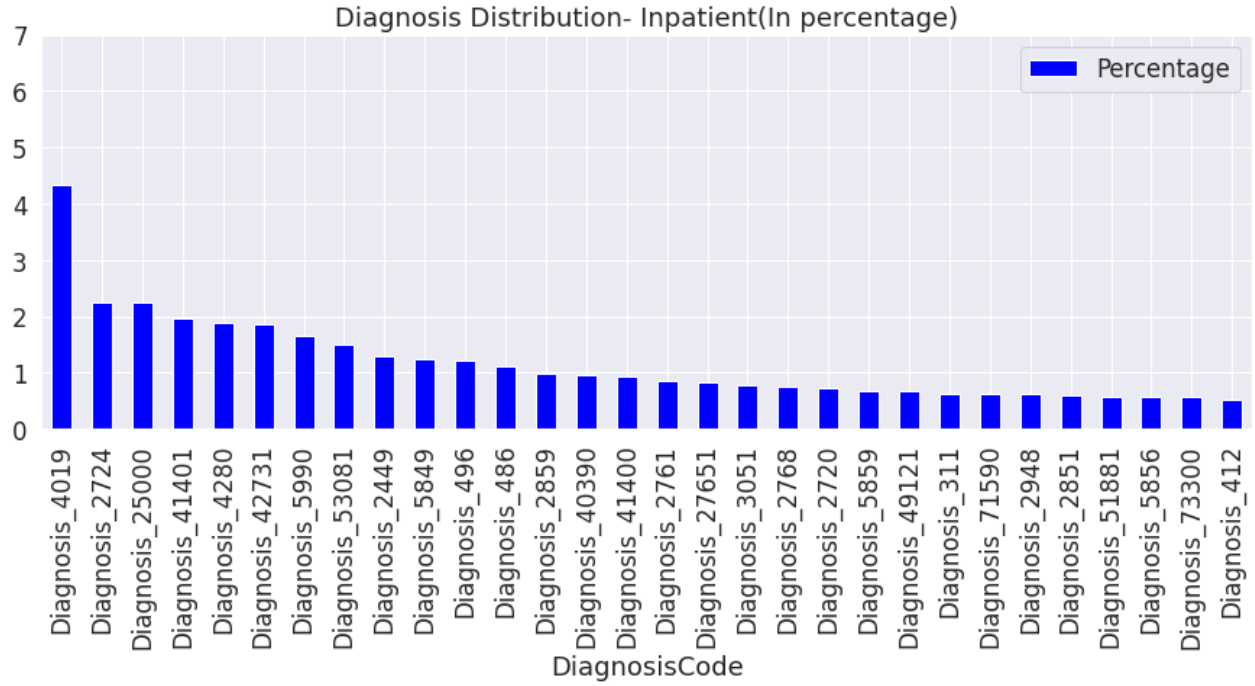

Figure 41: Most Common Diagnosis Codes Used by Fraudulent Providers of Inpatient

In Figure 42 we observe that procedure code 9904 is the most common procedure that is followed, around 5.1% of outpatients have undergone procedure code 9904, and procedure codes 9904, 4516, 66, 9390, and 8703 are the top 5 procedure code for outpatient data.

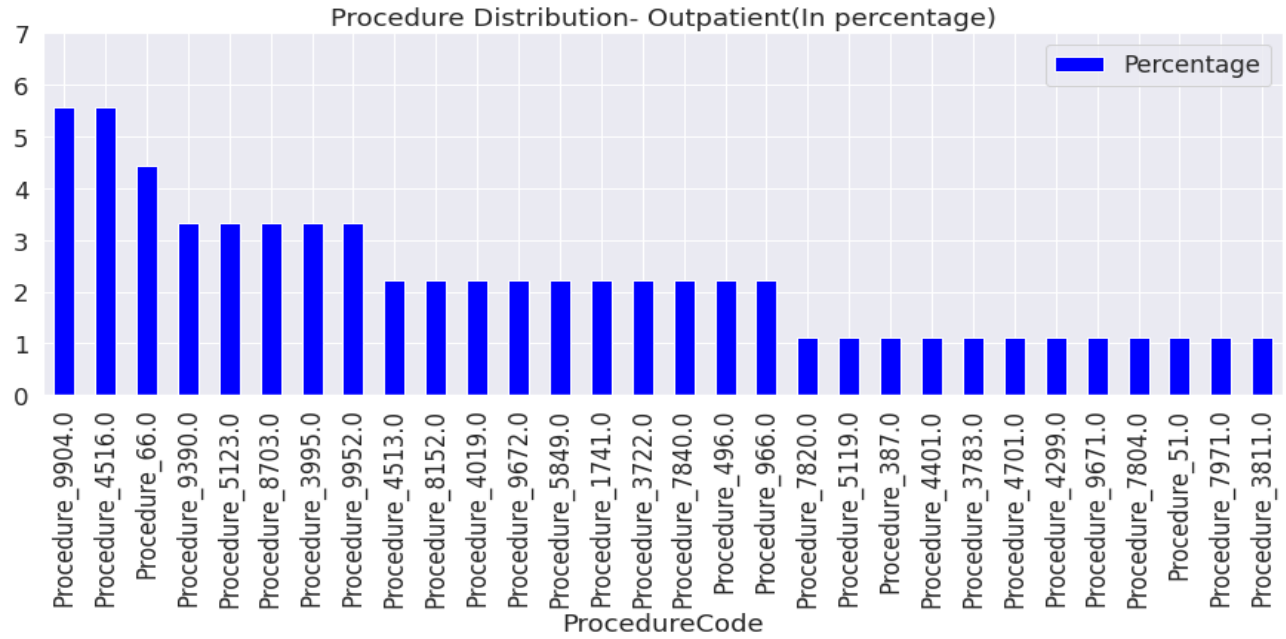

Figure 42: Most Common Procedure Codes Used by Fraudulent Providers of Outpatient

In Figure 43 we observe that diagnosis code 4019 is the most common diagnosis an outpatient undergoes, around 4.8% of outpatients have undergone diagnosis code 4019, and diagnosis codes 4019, 25000, 2724, V5869, and 401 are the top 5 Diagnosis codes for outpatient data.

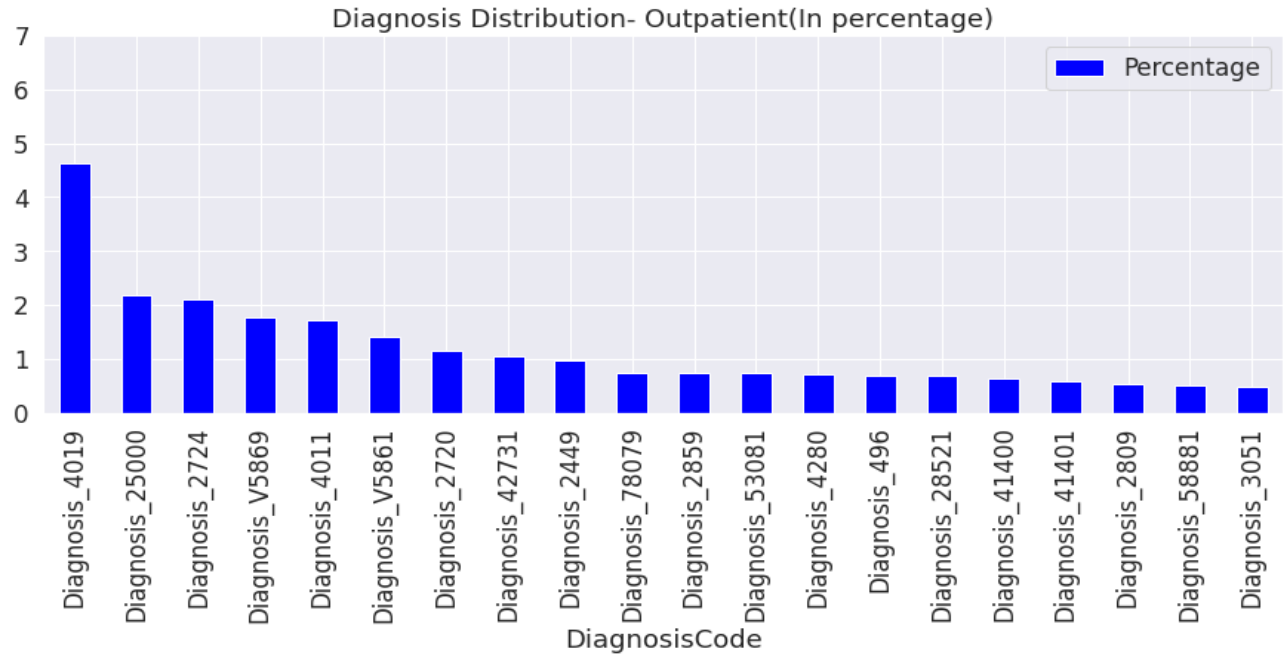

Figure 43: Most Common Diagnosis Codes Used by Fraudulent Providers of Outpatient

5.5.1.6.3 Most Common Codes of Physicians Used by Fraudulent Providers of Inpatient and Outpatient

In Figure 44 we observe that most inpatients are attended by physician PHY422134, and around 1% of the inpatients are attended by physician PHY422134.

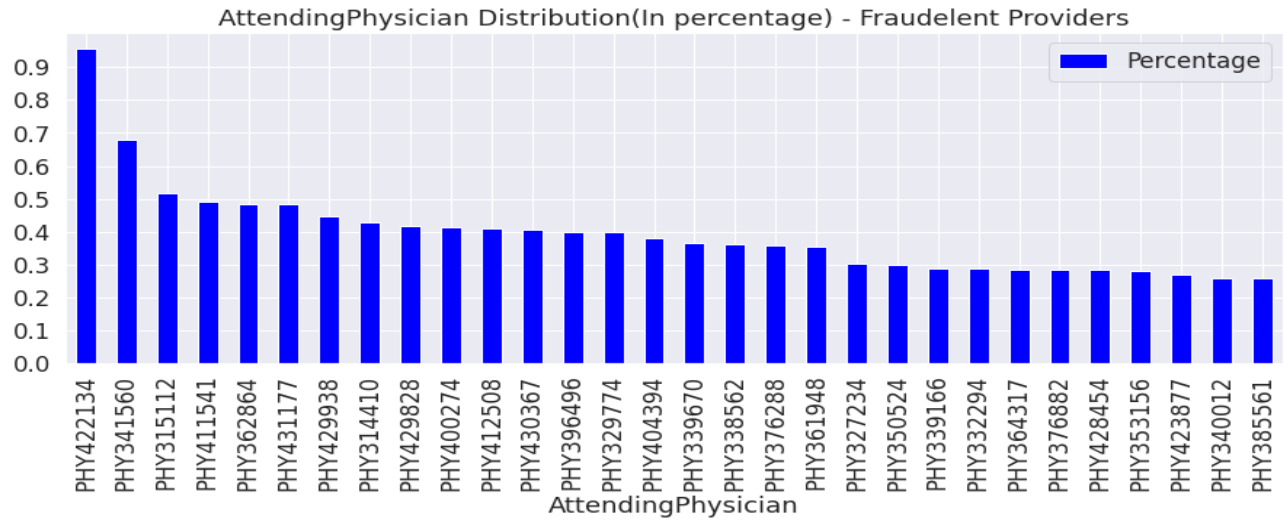

Figure 44: Most Common Codes of Attending Physicians Used by Fraudulent Providers of Inpatient

In Figure 45 we observe that physician PHY429430 performs most of the operations and around 1% of the inpatients are attended by physician PHY429430.

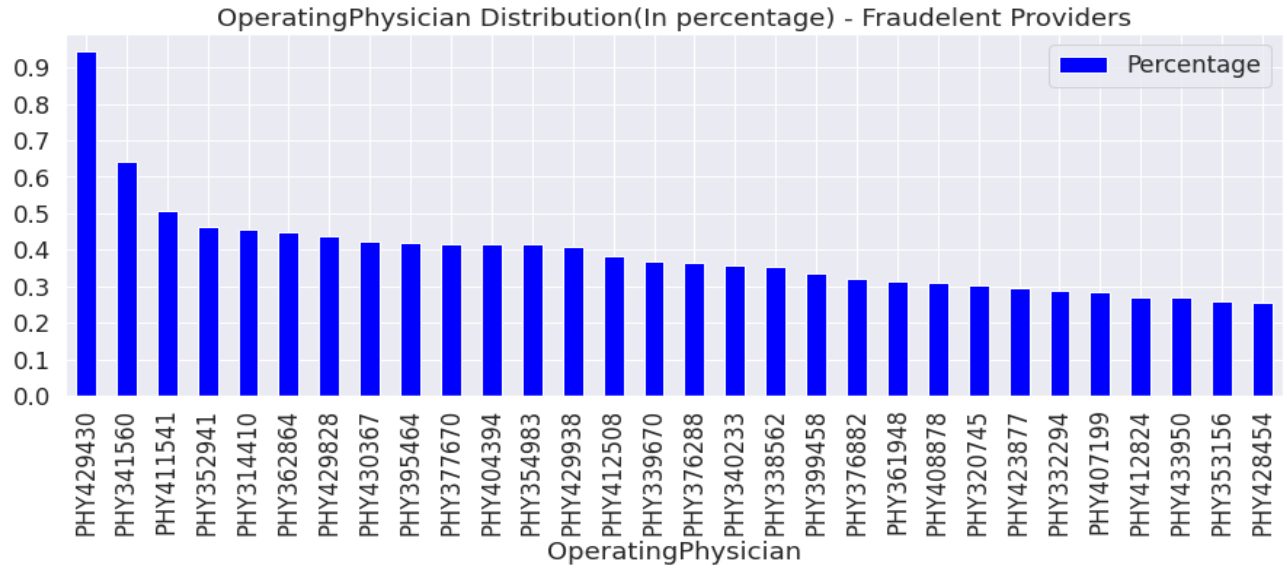

Figure 45: Most Common Codes of Operating Physicians Used by Fraudulent Providers of Inpatient

In Figure 46 we observe that the most common code of other physicians is PHY416093.

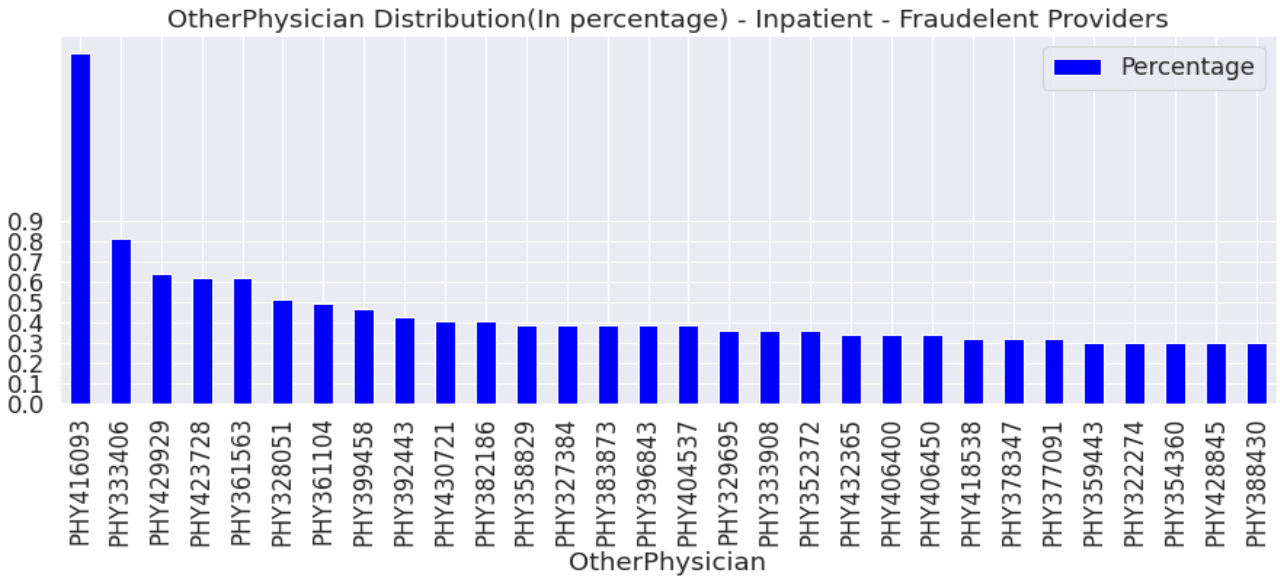

Figure 46: Most Common Codes of Other Physicians Used by Fraudulent Providers of Inpatient

In Figure 47 we observe that most patients are attended by physician PHY330576 and around 0.5% of the outpatients are attended by physician PHY330576.

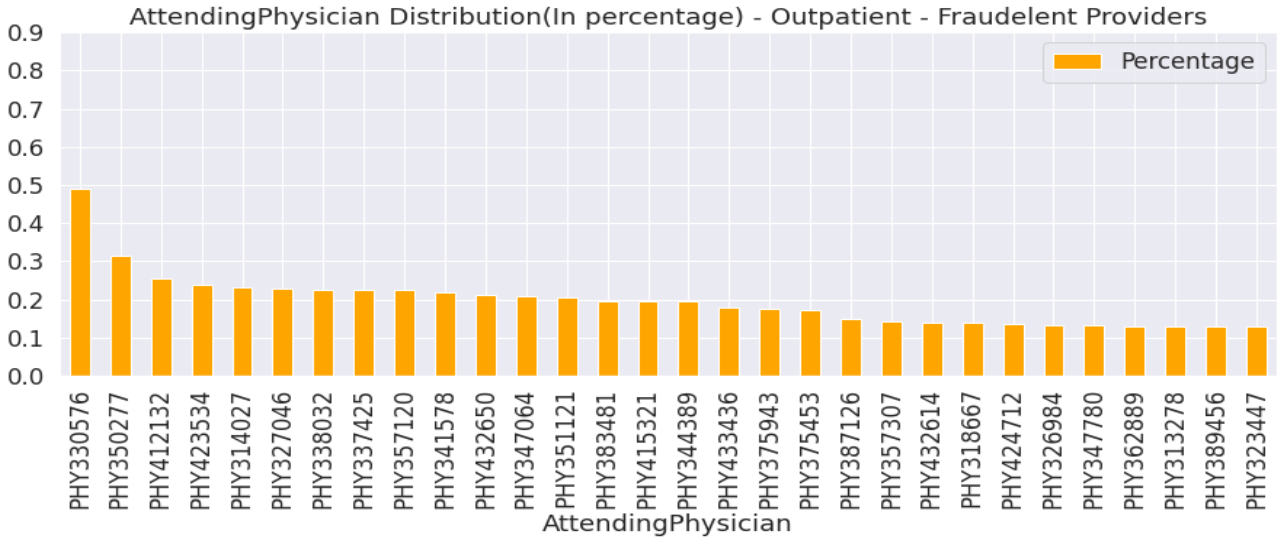

Figure 47: Most Common Codes of Attending Physicians Used by Fraudulent Providers of Outpatient

In Figure 48 we observe that physician PHY330576 performs most of the operations and around 0.48% of the outpatients are attended by physician PHY330576.

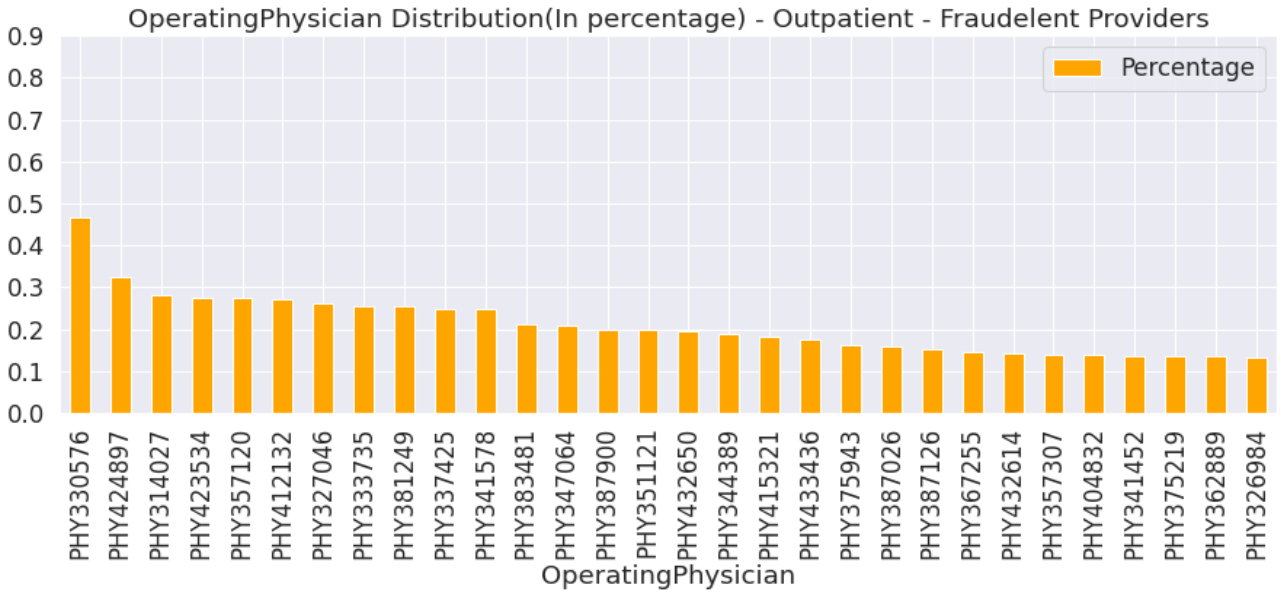

Figure 48: Most Common Codes of Operating Physicians Used by Fraudulent Providers of Outpatient

In Figure 49 we observe that the most common code of other physicians is PHY412132.

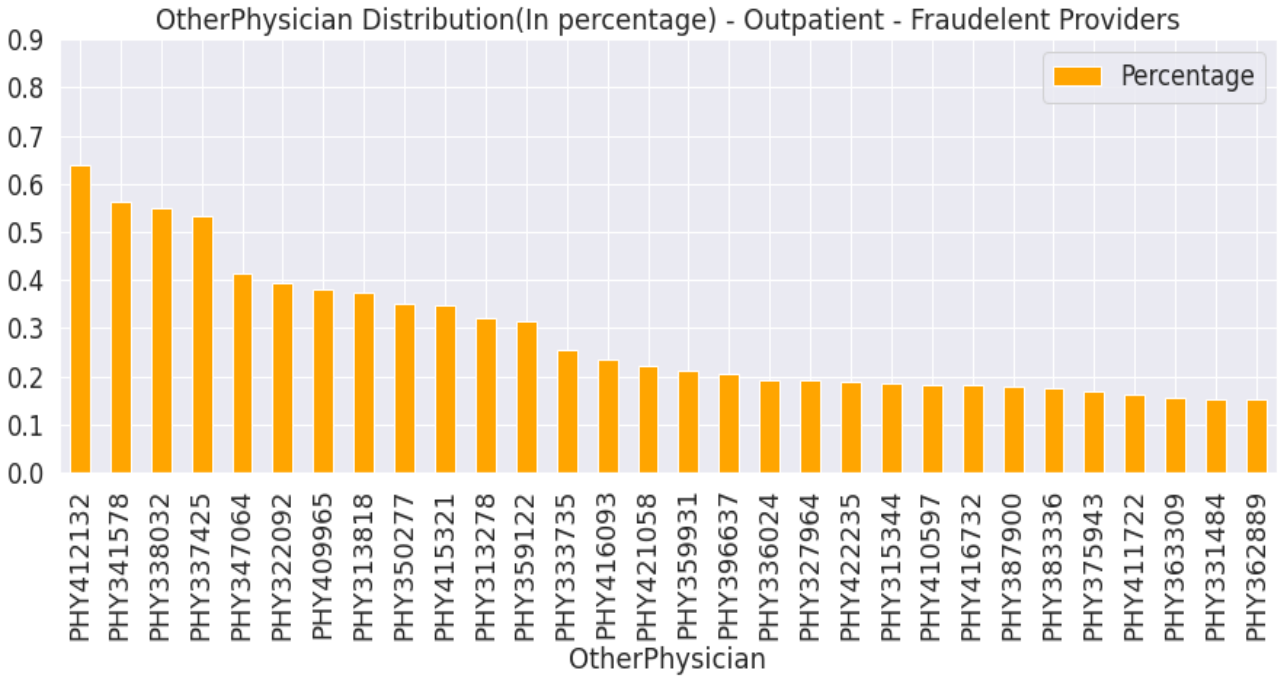

Figure 49: Most Common Codes of Other Physicians Used by Fraudulent Providers of Outpatient

5.5.1.6.4 Most Common States, Countries, and Race Used by Fraudulent Providers of Inpatient and Outpatient

In Figures 50, 51, and 52 we observe that states coded as 5, 10, 33, and 45 have the most fraudulent encounters for Inpatient data. County coded as 10, 200, 160, and 60 have the most fraudulent encounters for Inpatient data. Race 1 has the most fraudulent encounters for Inpatient data with 84.33%.

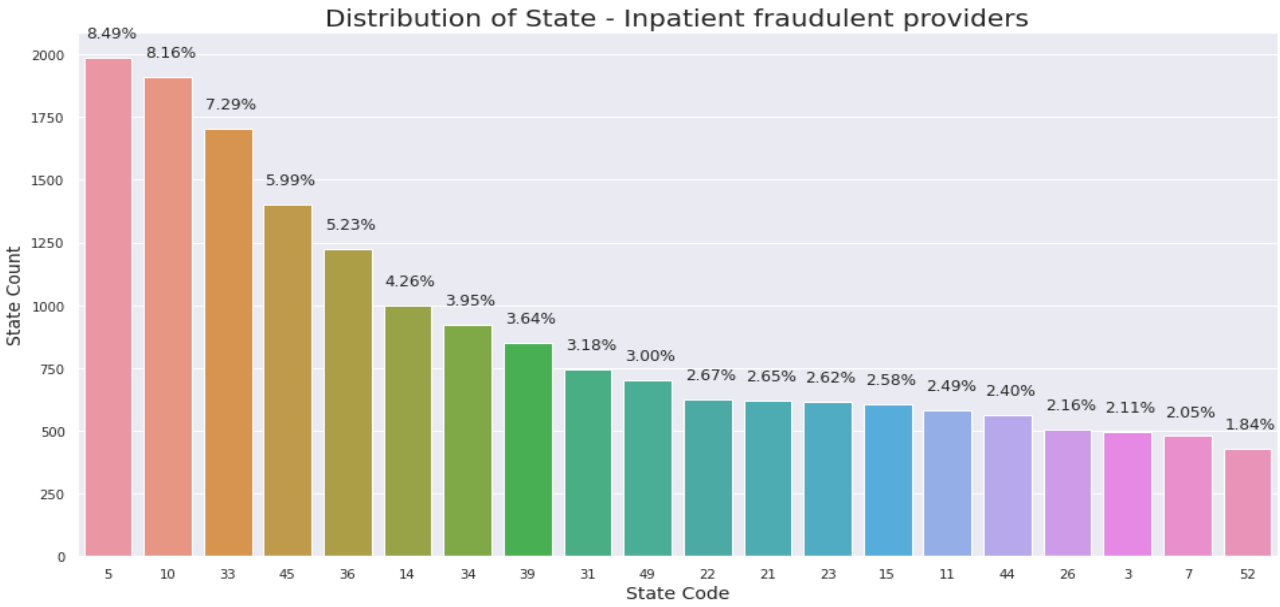

Figure 50: Most Common States of Inpatient of Fraudulent Providers

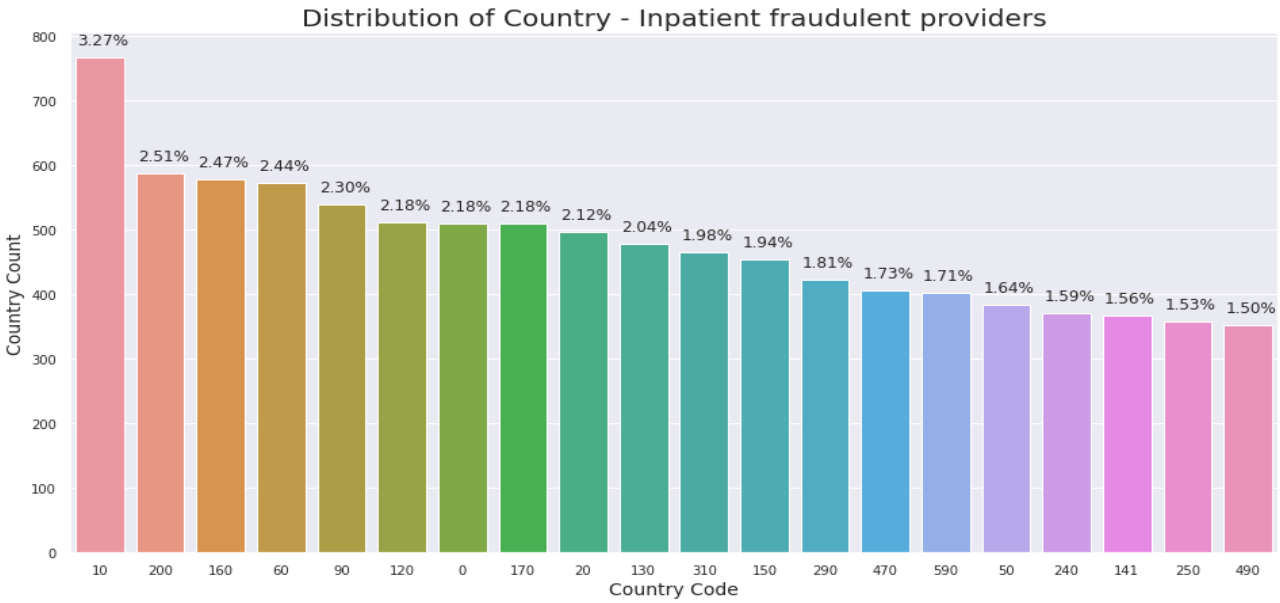

Figure 51: Most Common Countries of Inpatient Used by Fraudulent Providers

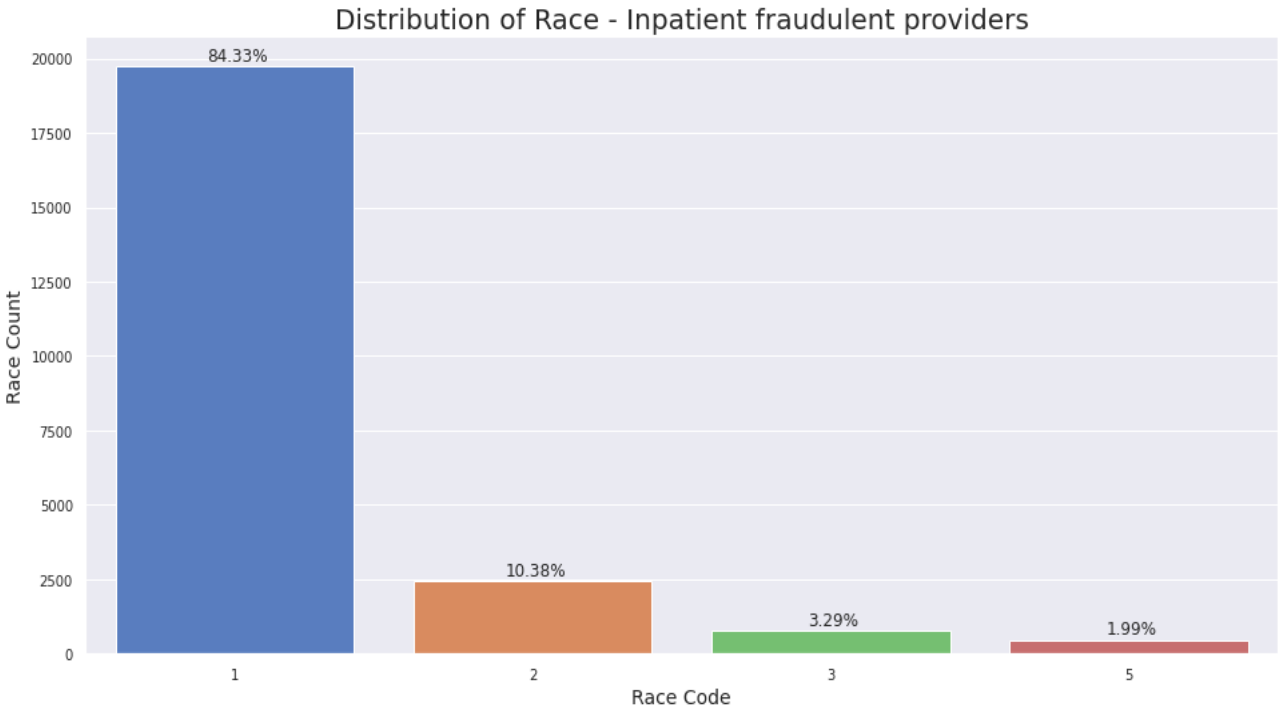

Figure 52: Most Common Races of Inpatients Used by Fraudulent Providers

In Figures 53, 54, and 55 we observe that states coded as 5, 33 10, and 39 have the most fraudulent encounters for outpatient data. County coded as 200, 470, 400, and 590 have the most fraudulent encounters for inpatient data. Race 1 has the most fraudulent encounters for inpatient data with 83.75%.

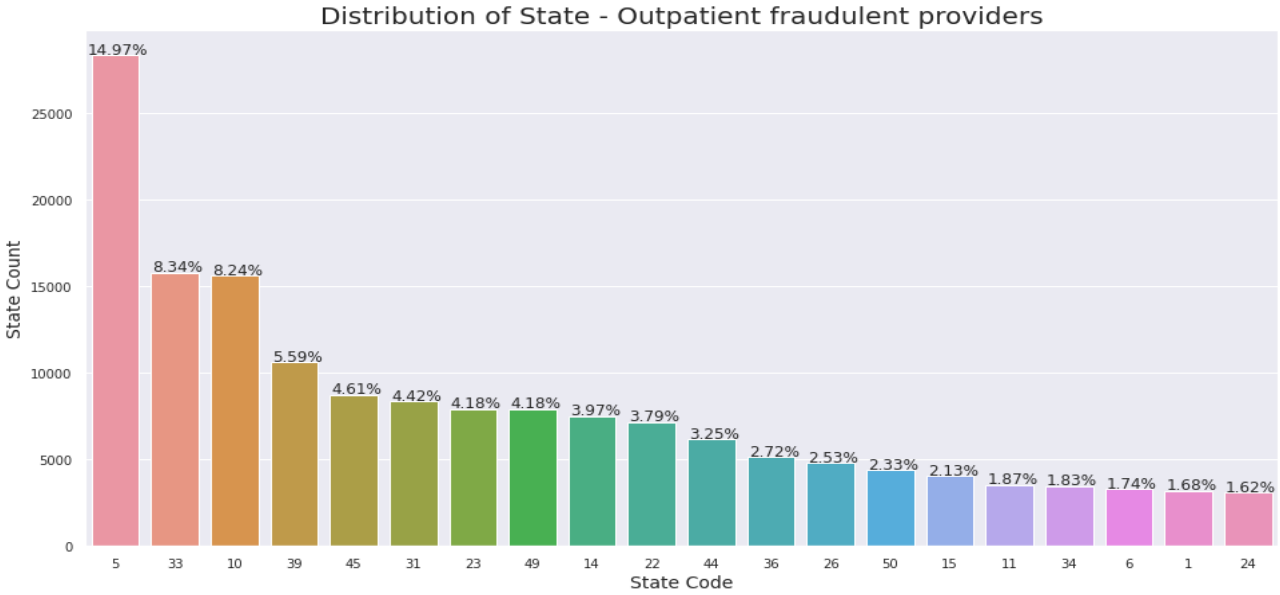

Figure 53: Most Common States of Outpatients Used by Fraudulent Providers

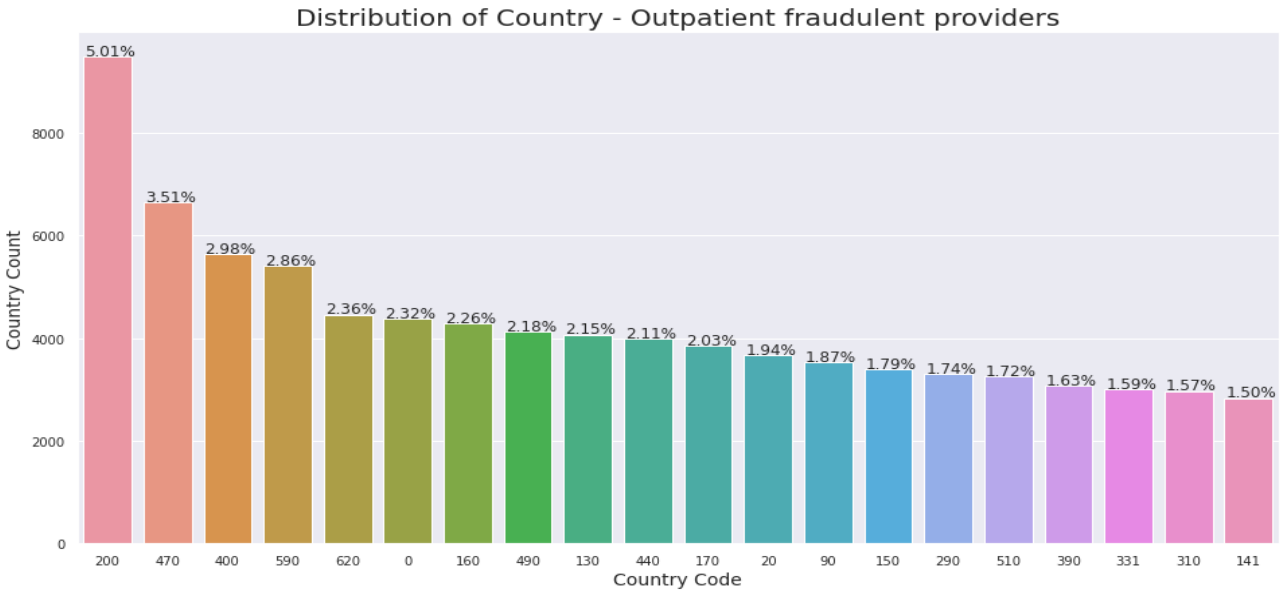

Figure 54: Most Common Countries of Outpatients Used by Fraudulent Providers

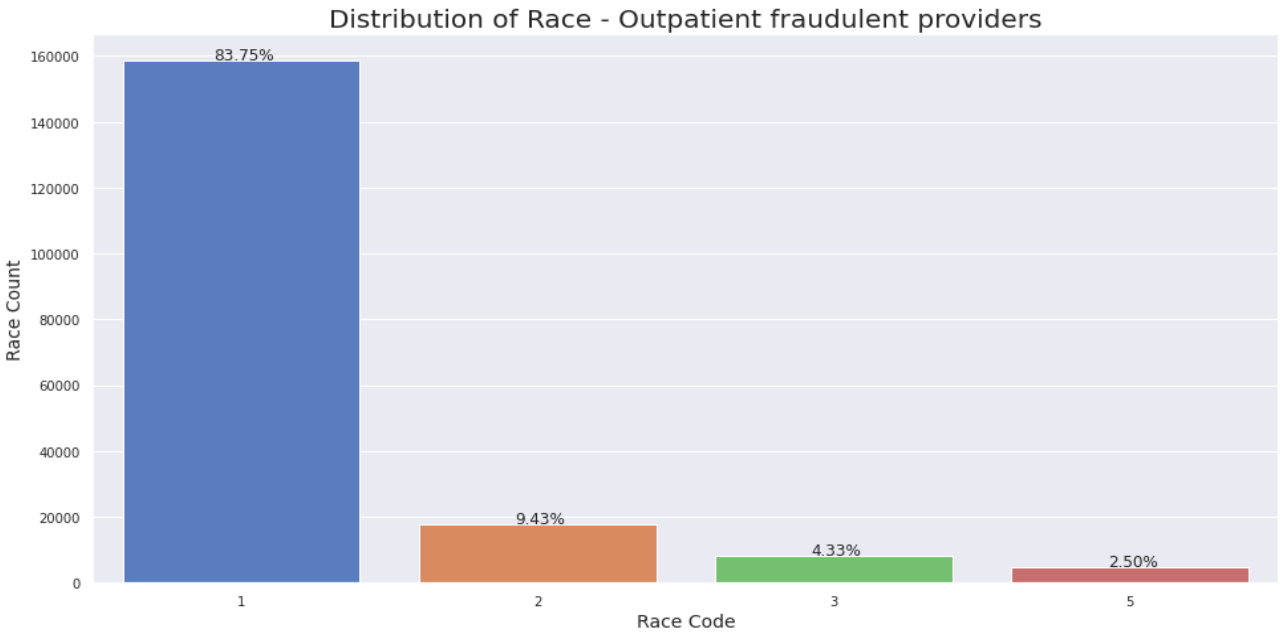

Figure 55: Most Common Races of Outpatients Used by Fraudulent Providers

5.5.1.6.5 Date of Birth of Inpatient and Outpatient Used by Fraudulent Providers

In Figure 56 we observe that most fraudulent encounters in inpatient data are observed for the patients born between 1920 and 1945.

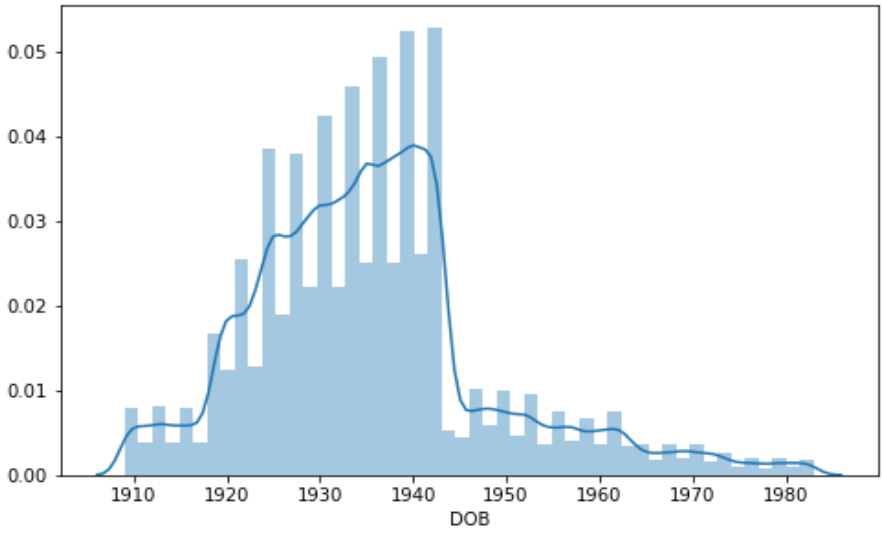

Figure 56: Date of Birth of Inpatient and Outpatient Used by Fraudulent Providers

In Figure 57 we observe that most fraudulent encounters in outpatient data are observed for the patients born between 1920 and 1945.

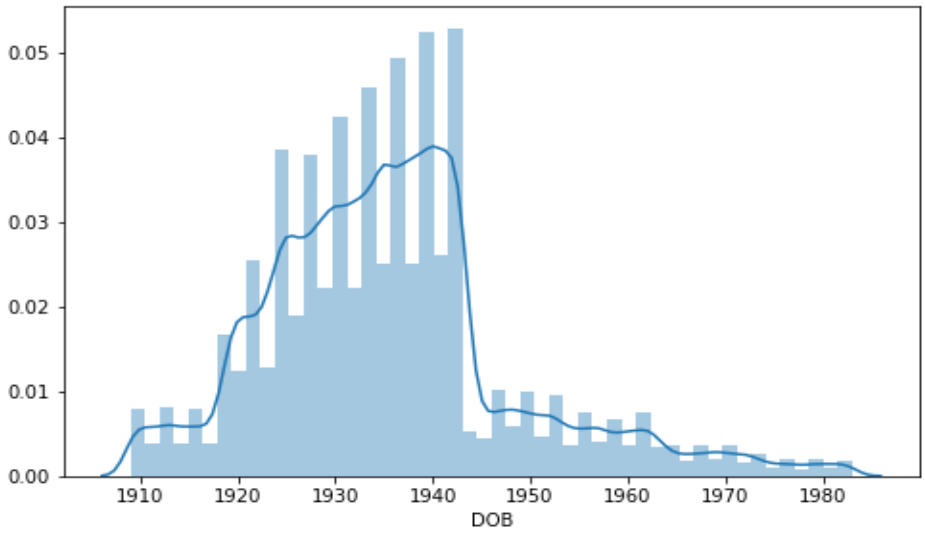

Figure 57: Date of Birth of Outpatient Used by Fraudulent Providers

5.5.1.6.6 Money Lost in Fraud by Fraudulent Providers

In Figure 58 we observe that the distribution of the amount that is paid as claim reimbursement seems like a log-normal distribution. Most of all reimbursed amount is between 0 and 20000 and in very few cases amount more than 20000 is paid for claim reimbursement. Total amount of reimbursement for inpatient is 241,288,510 and for outpatient is 54,392,610. So, the total money for reimbursements as per the data are 295,543,140. That is around 290 million.

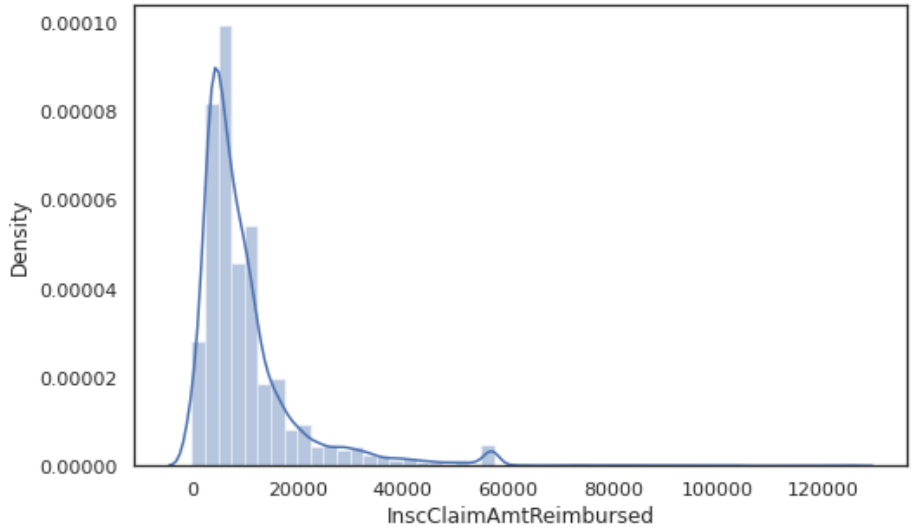

Figure 58: Money Lost in Fraud by Fraudulent Providers of Inpatient

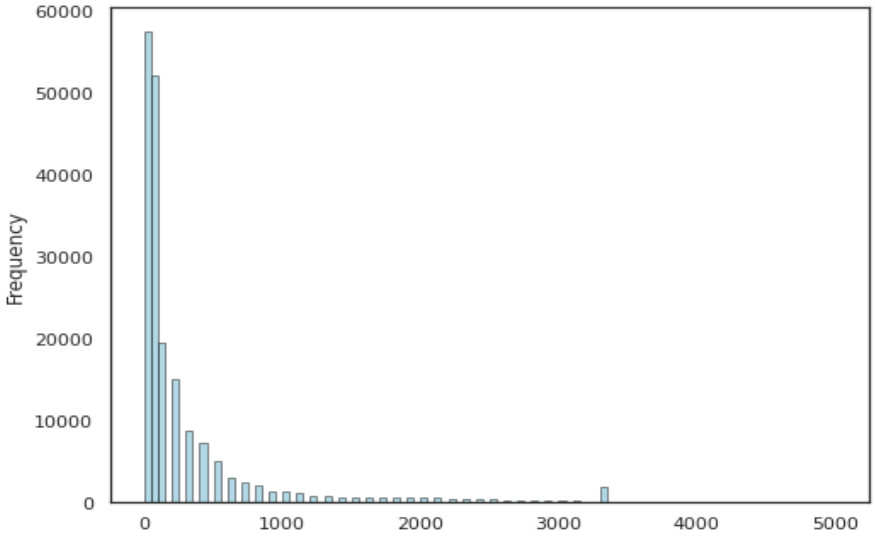

Figure 59: Money Lost in Fraud by Fraudulent Providers of Outpatient

5.5.2 Merged Data

5.5.2.1 Distribution of Class Data

Figure 60 shows that after merging the datasets there are 38.12% fraudulent providers and 61.88% non-fraudulent providers.

Figure 60: Distribution of Class Labels - Merge Data

5.5.2.2 Most Common Procedure and Diagnosis Codes

In Figure 61 we observe that procedure code 4019 is the most common procedure that is followed, around 6.6% of patients have undergone procedure code 4019, and procedure codes 4019, 9904, 2724, 8154, and 66 are the top 5 procedure codes for merge data.

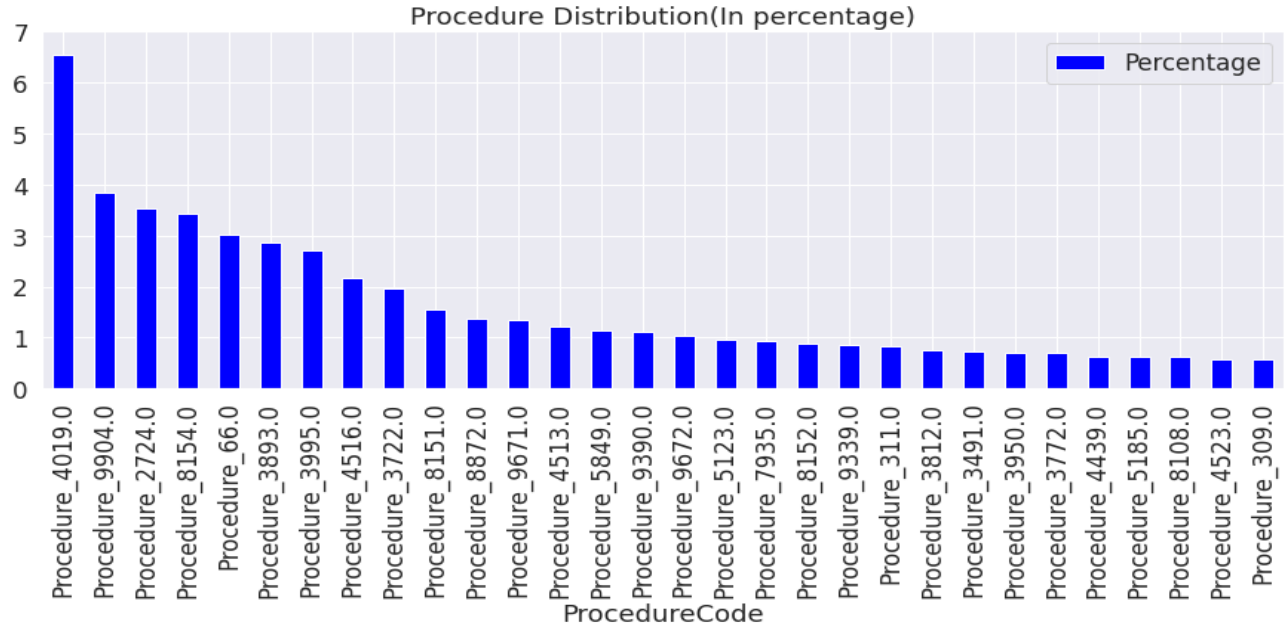

Figure 61: Most Common Procedure Codes - Merge Data

In Figure 62 we observe that diagnosis code 4019 is the most common diagnosis a patient undergoes, around 4.5% of patients have undergone diagnosis code 4019, and diagnosis codes 4019, 25000, 2724, 5869, and 4011 are the top 5 diagnosis codes for merge data.

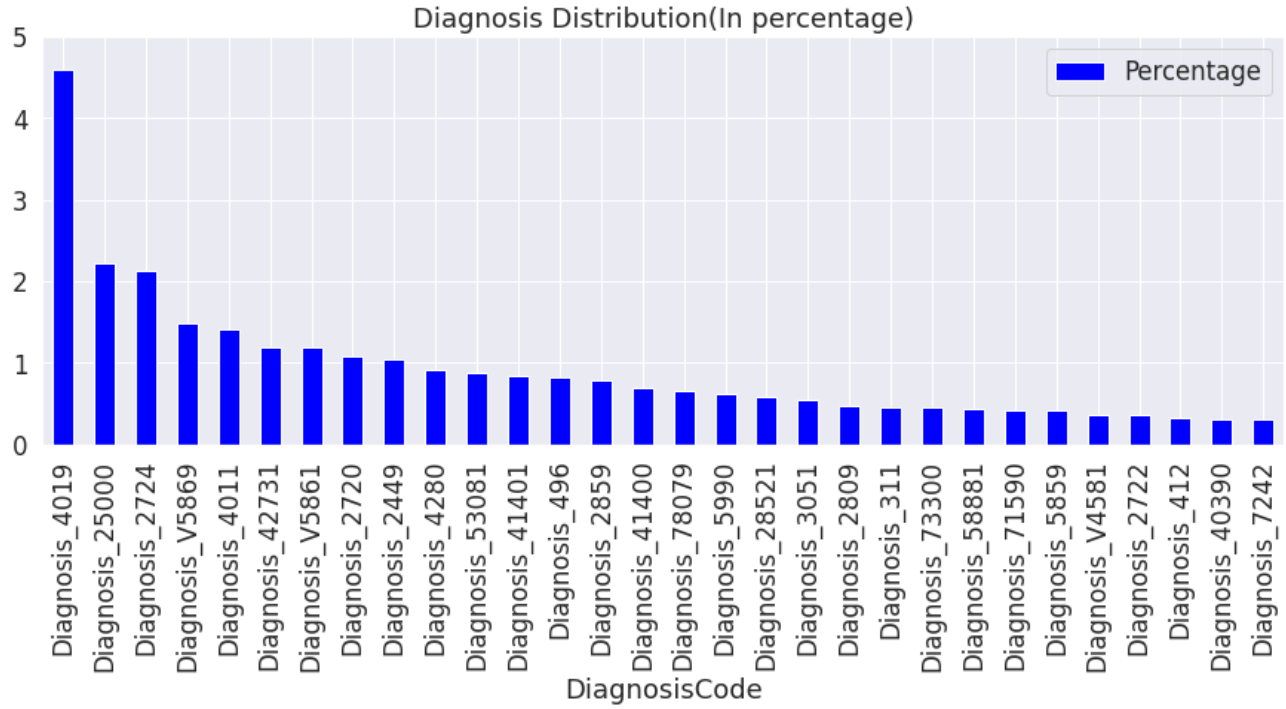

Figure 62: Most Common Diagnosis Codes - Merge Data

5.5.2.3 Most Common Codes of Physicians

In Figure 63 we observe that most patients are attended by physician PHY330576 and around 0.45% of the patients are attended by physician PHY330576.

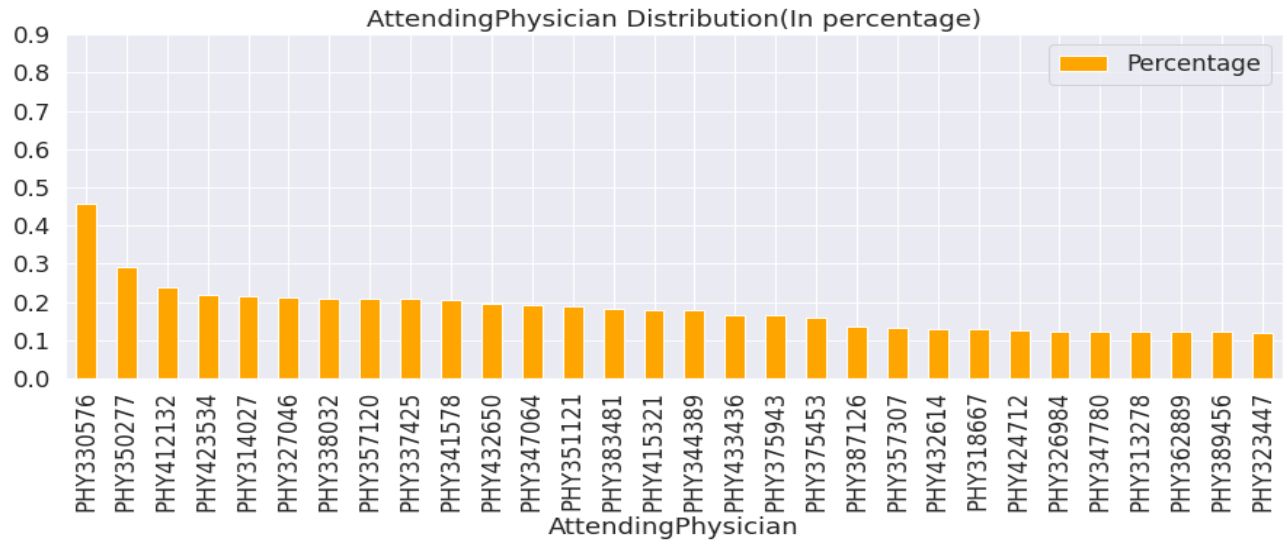

Figure 63: Most Common Codes of Attending Physicians - Merge Data

In Figure 64 we observe that most patients are attended by physician PHY330576 and around 0.39% of the patients are attended by physician PHY330576.

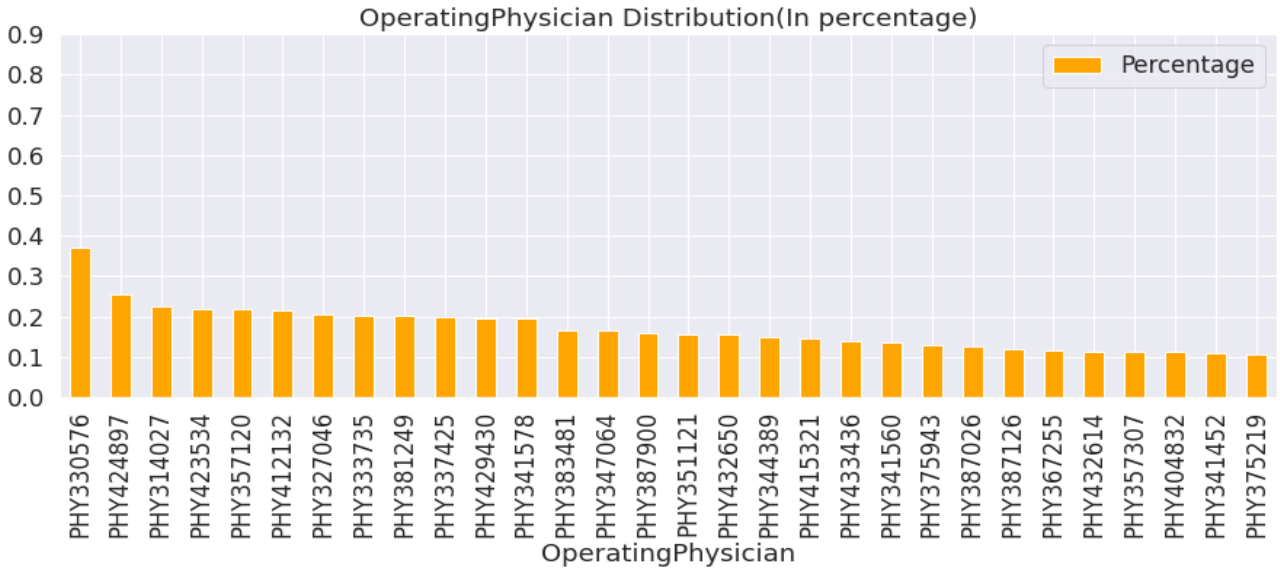

Figure 64: Most Common Codes of Operating Physicians - Merge Data

In Figure 65 we observe that most patients are attended by physician PHY412132 and around 0.62% of the patients are attended by physician PHY412132.

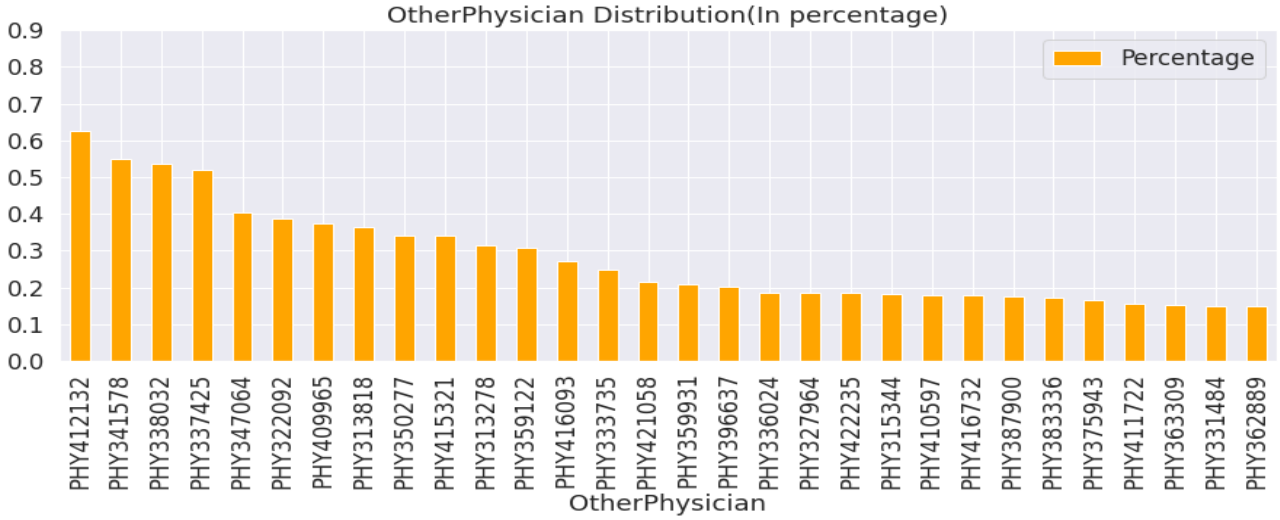

Figure 65: Most Common Codes of Other Physicians - Merge Data

5.5.2.4 Most Common States, Countries, and Races

In Figures 66, 67, and 68 we observe that states coded as 5, 10, 33, and 45 are the most common a patient outcomes from for merge data. County coded as 200, 10, 20, and 470 are the most common a patient outcomes from for merge data. Race 1 is the most common a patient outcomes from for merge data with 84.38%.

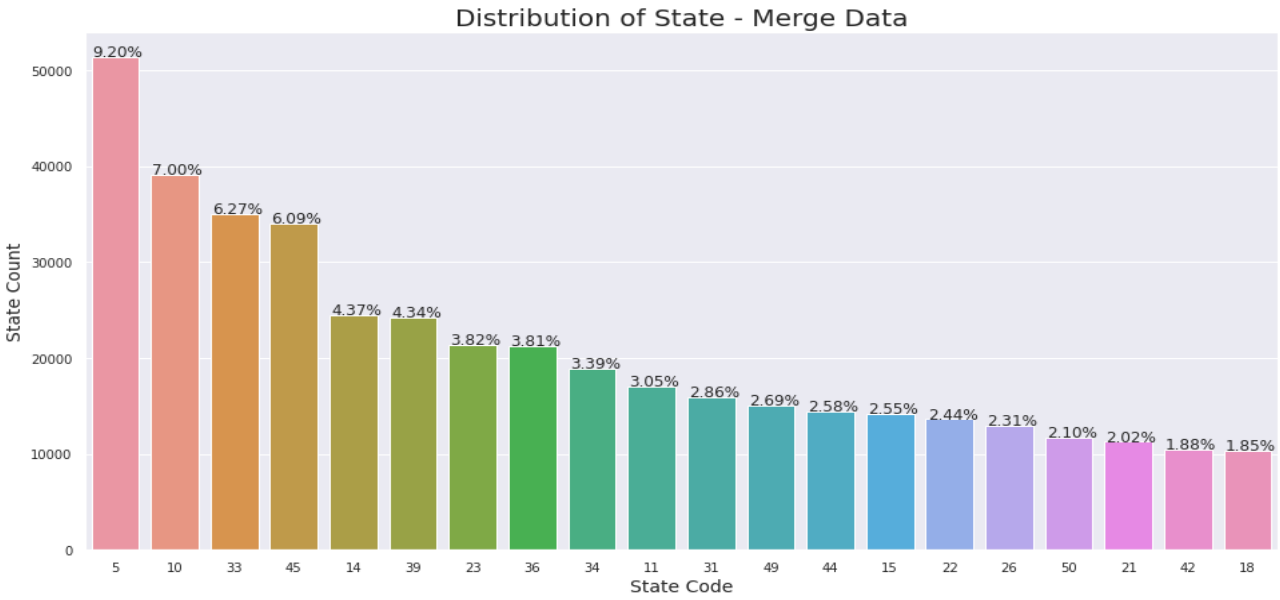

Figure 66: Most Common States of Inpatient of Fraudulent Providers - Merge Data

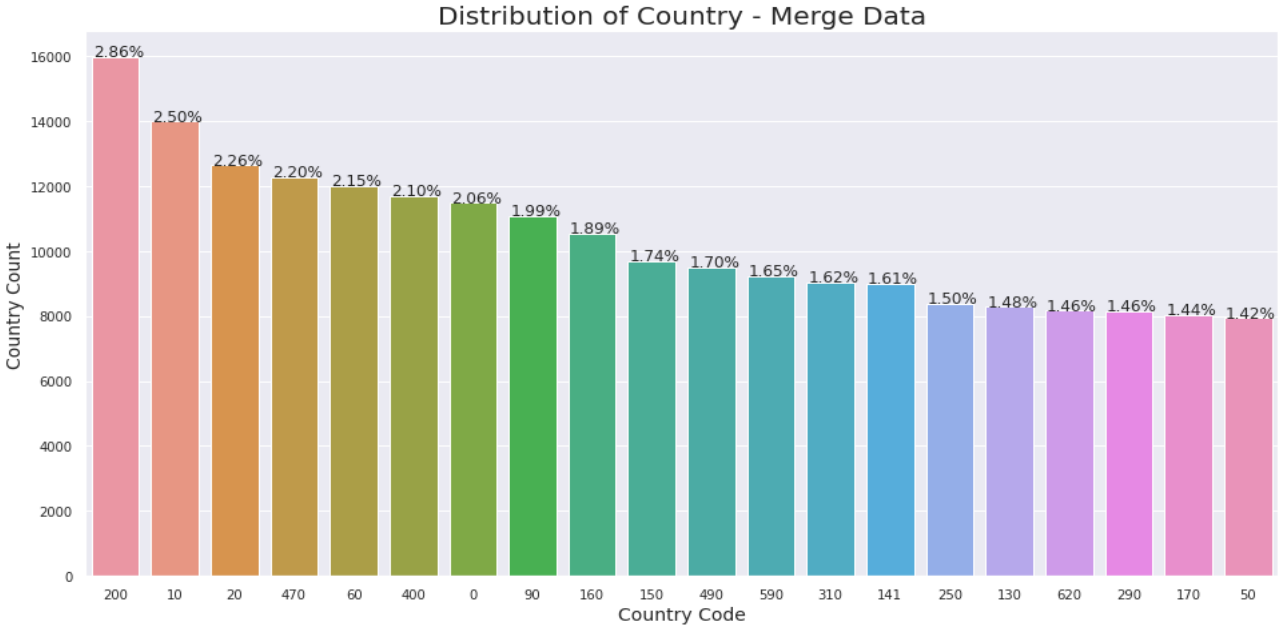

Figure 67: Most Common Countries of Inpatient by Fraudulent Providers - Merge Data

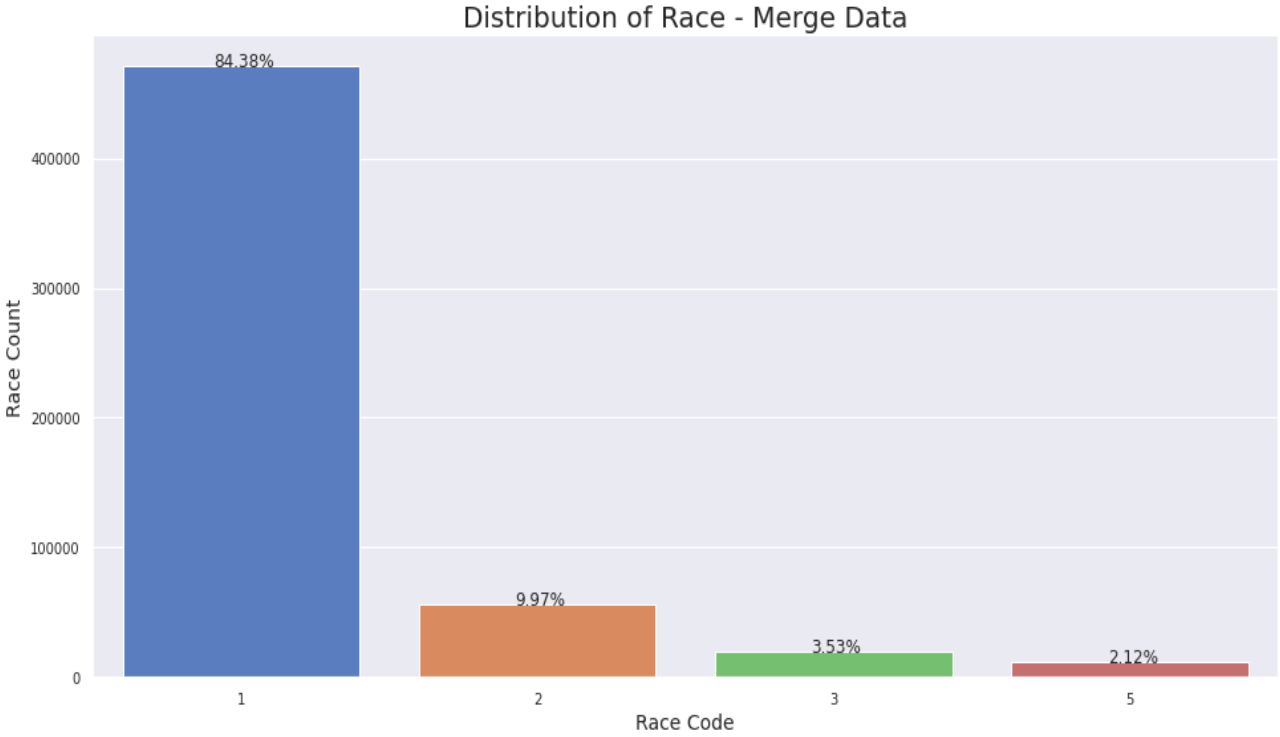

Figure 68: Most Common Races of Inpatients by Fraudulent Providers - Merge Data

5.5.2.5 Amount of Reimbursement

In Figure 69 we observe that the distribution of the amount that is paid as claim reimbursement seems like a log-normal distribution. Most of all reimbursed amount is between 0 and 20000 and in very few cases amount more than 20000 is paid for claim reimbursement. The total amount of reimbursement for inpatients is 556,543,140. That is around 556 million.

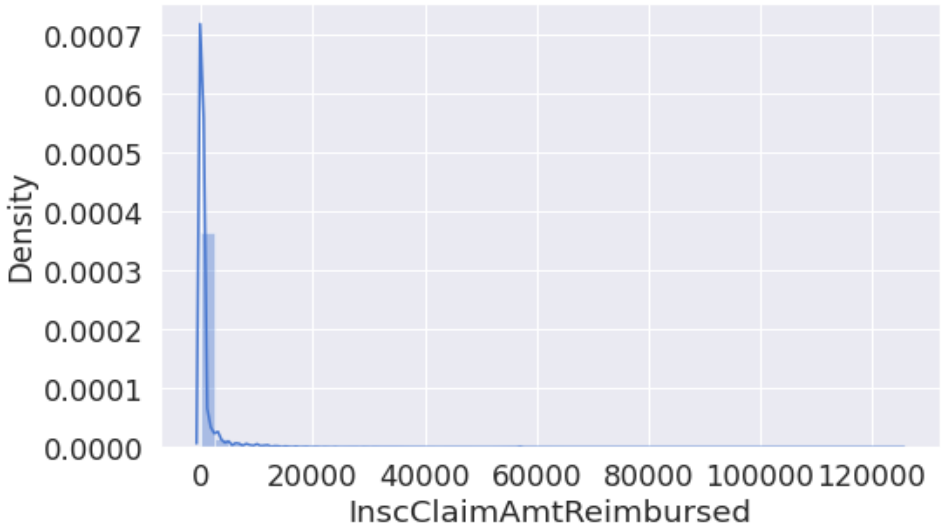

Figure 69: Reimbursement Amount - Merge Data

5.5.2.6 Fraudulent Providers

5.5.2.6.1 Percentage of Fraudulent Providers of Inpatient and Outpatient

Figure 70 shows that there are 10.12% of fraudulent providers in inpatients and 89.9% in outpatients.

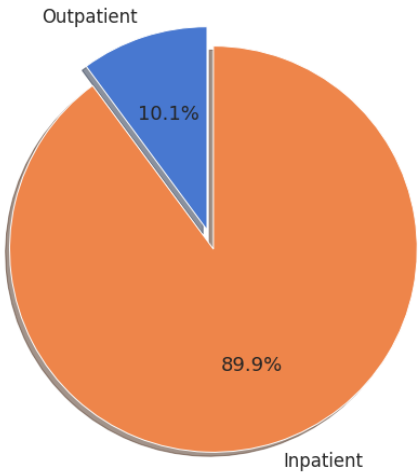

Figure 70: Percentages of Inpatient and Outpatient for Fraudulent Providers

5.5.2.6.2 Most Common Procedure and Diagnosis Codes Used by Fraudulent Providers

In Figure 71 we observe that procedure code 4019 is the most common procedure that is followed, around 6.6% of patients have undergone procedure code 4019, and procedure codes 4019, 2724, 9904, and 8154 are the top 5 procedure codes for fraudulent providers in merge data.

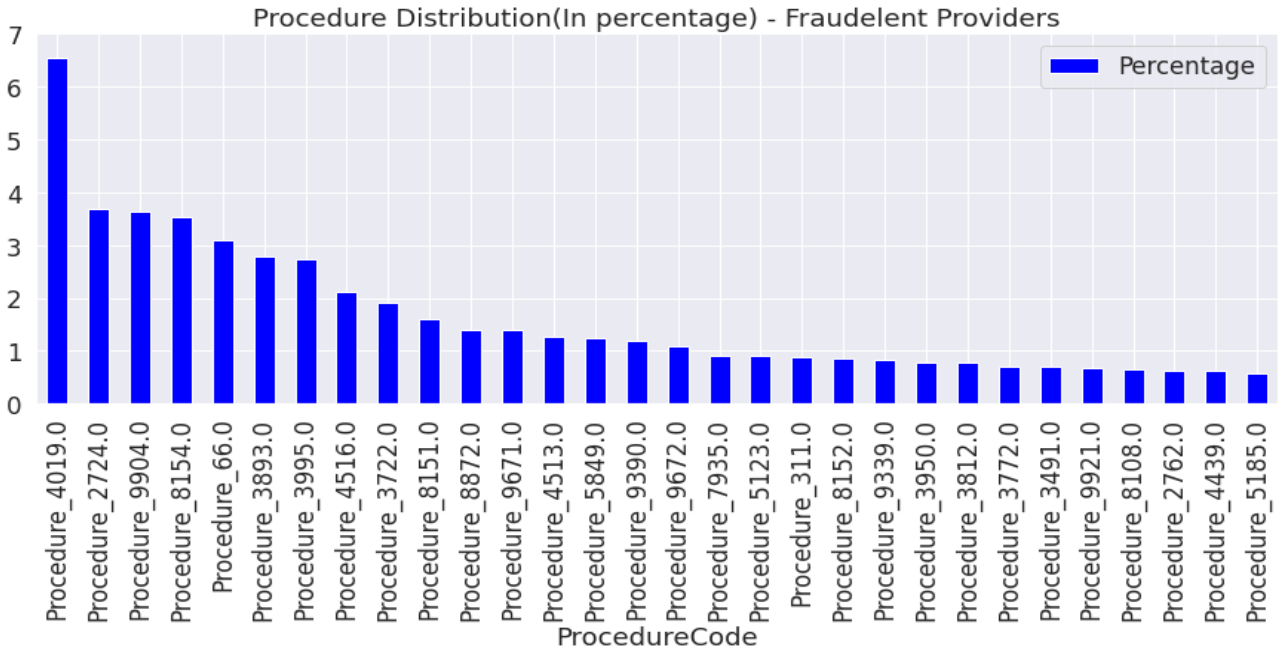

Figure 71: Most Common Procedure Codes by Fraudulent Providers - Merge Data

In Figure 72 we observe that diagnosis code 4019 is the most common diagnosis a patient undergoes, around 4.5% of patients have undergone diagnosis code 4019, and diagnosis codes 4019, 25000, 2724, 5869, and 42731 are the top 5 diagnosis codes for merge data.

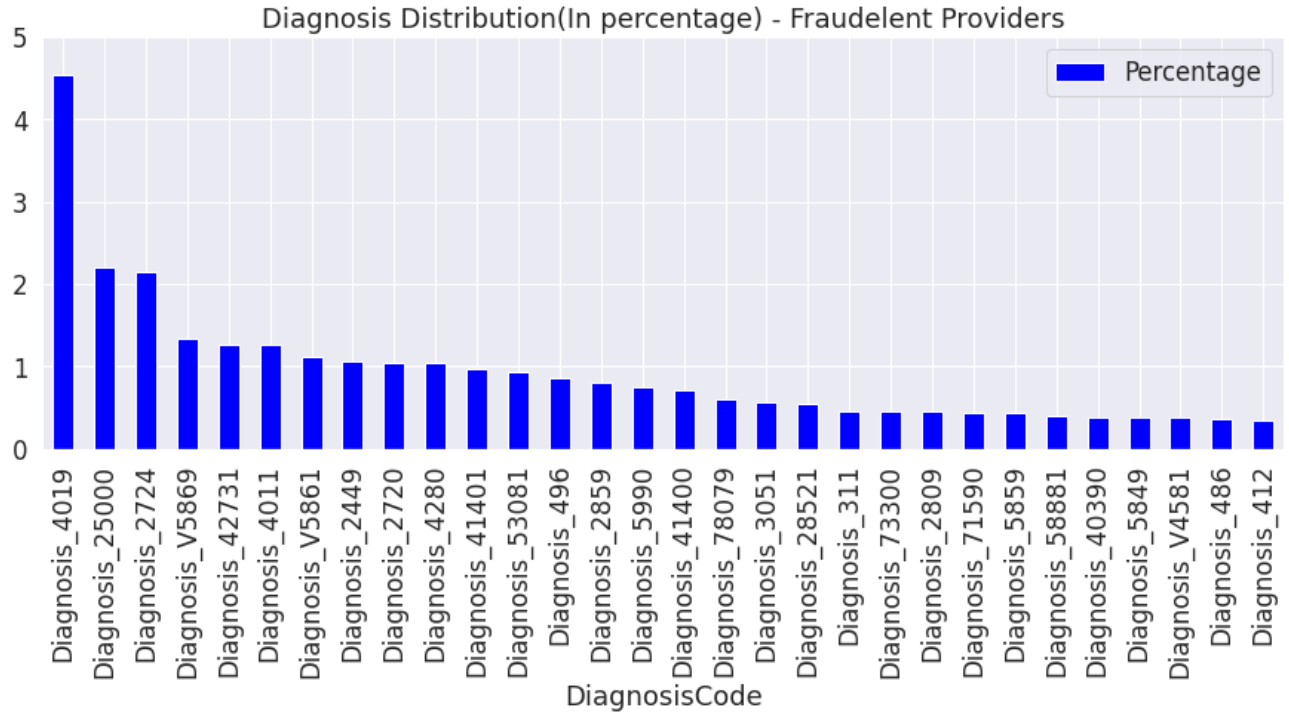

Figure 72: Most Common Diagnosis codes by Fraudulent Providers - Merge Data

5.5.2.6.3 Most Common Codes of Physicians Used by Fraudulent Providers

In Figure 73 we observe that most patients are attended by a physician PHY330576 and around 1% of the patients are attended by a physician PHY330576.

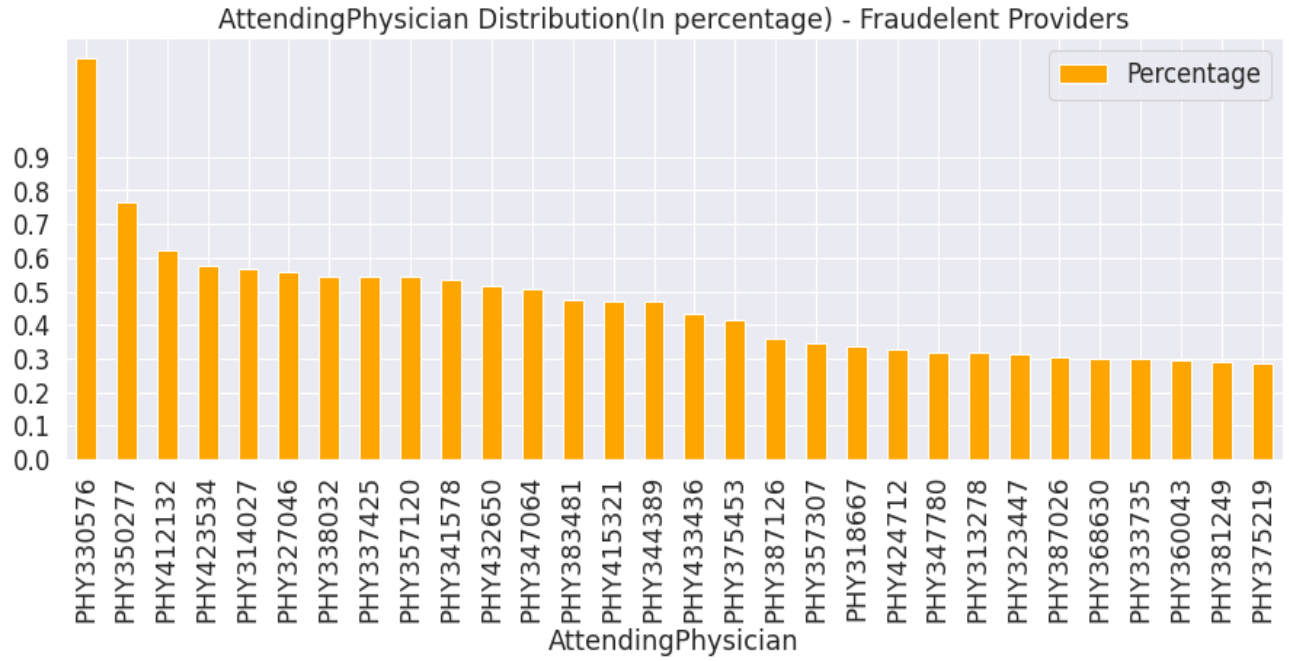

Figure 73: Most Common Codes of Attending Physicians Used by Fraudulent Providers - Merge Data

In Figure 74 we observe that most patients are attended by physician PHY330576 and around 0.9% of the patients are attended by physician PHY330576.

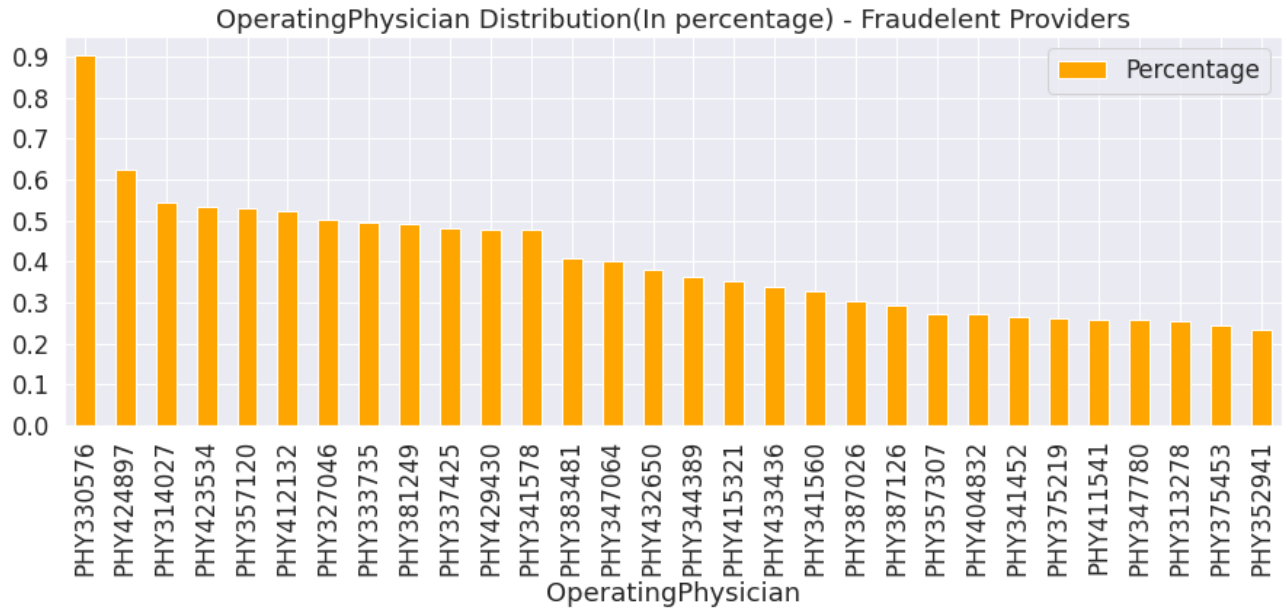

Figure 74: Most Common Codes of Operating Physicians by Fraudulent Providers - Merge Data

In Figure 75 we observe that most patients are attended by physician PHY412132.

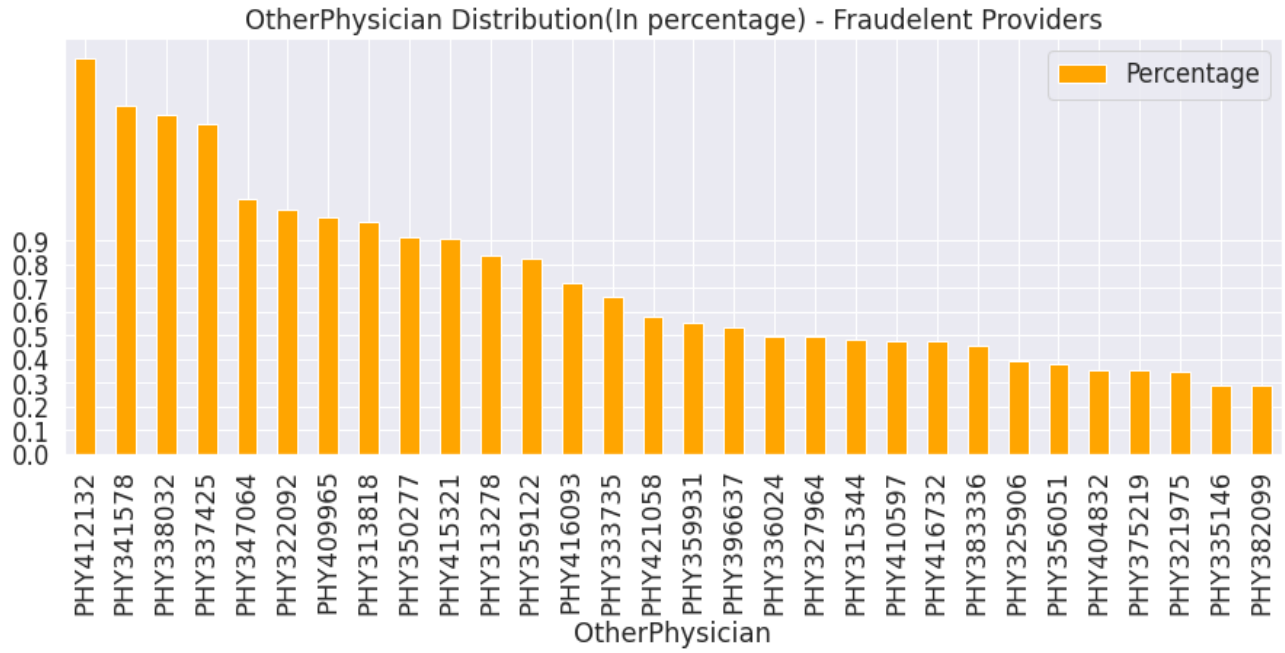

Figure 75: Most Common Codes of Other Physicians by Fraudulent Providers - Merge Data

5.5.2.6.4 Most Common States, Countries, and Race Used by the Potential Fraudulent Providers

In Figures 76, 77, and 78 we observe that states coded as 5, 10, 33, and 45 have the most fraudulent encounters for Inpatient data. County coded as 10, 200, 160, and 60 have the most fraudulent encounters for Inpatient data. Race 1 has the most fraudulent encounters for Inpatient data with 84.33%.

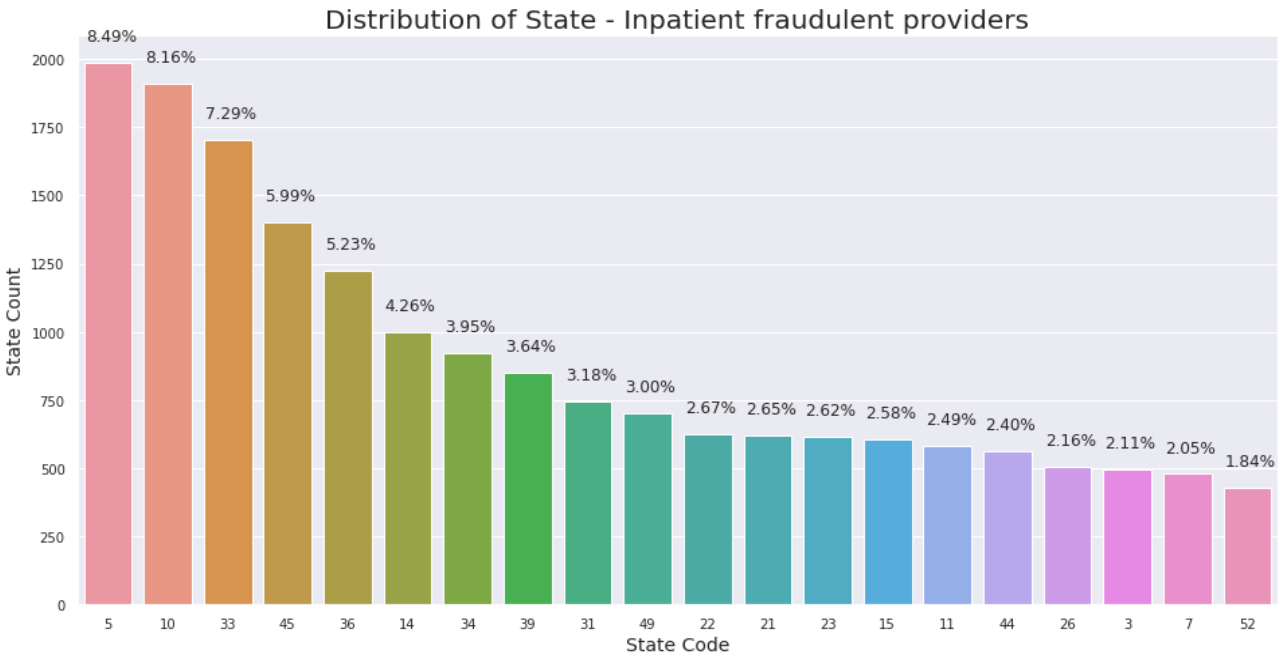

Figure 76: Most Common States of Fraudulent Providers - Merge Data

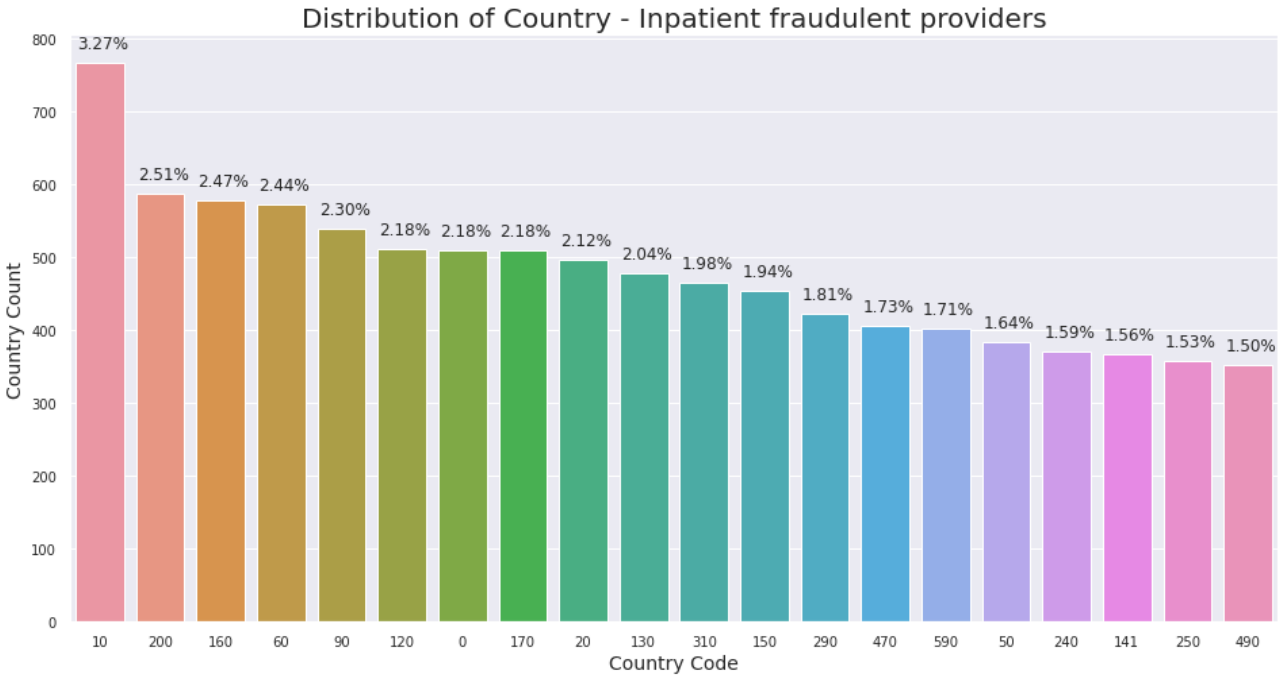

Figure 77: Most Common Countries of Fraudulent Providers - Merge Data

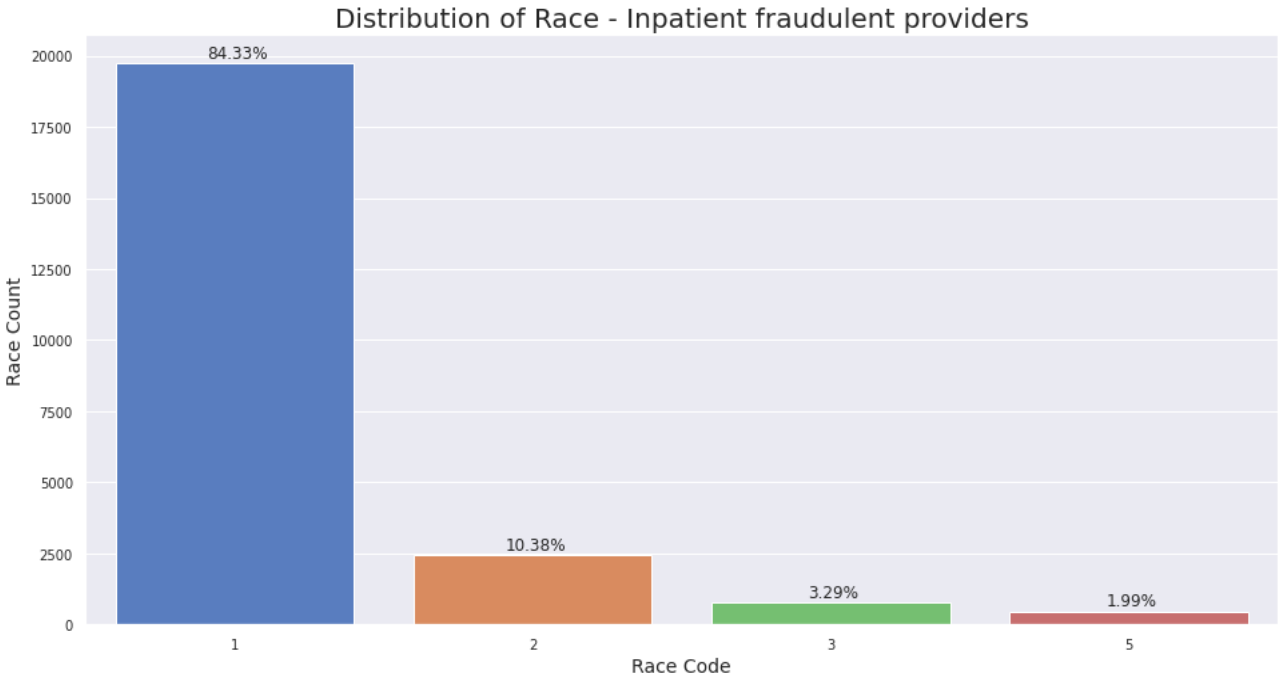

Figure 78: Most Common States of Fraudulent Providers - Merge Data

5.5.2.6.5 Date of Birth of Beneficiaries Used by Fraudulent Providers

In Figure 79 we observe that most fraudulent encounters are observed for the patients born between 1920 and 1945.

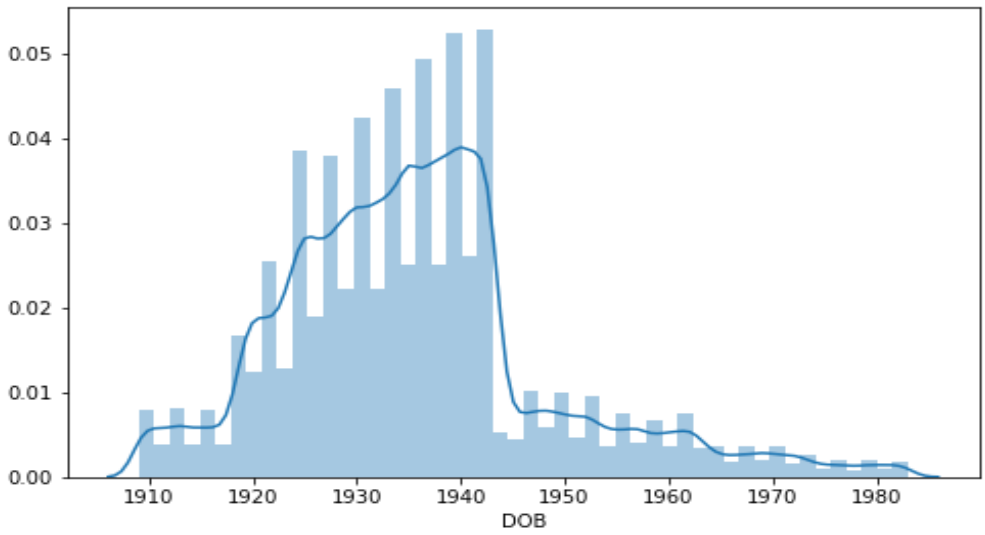

Figure 79: Date of Birth by Fraudulent Providers – Merge Data

5.5.2.6.6 Money Lost in Fraud by Fraudulent Providers

Total money lost as per merged data for 295,681,120. That is around 295 million.

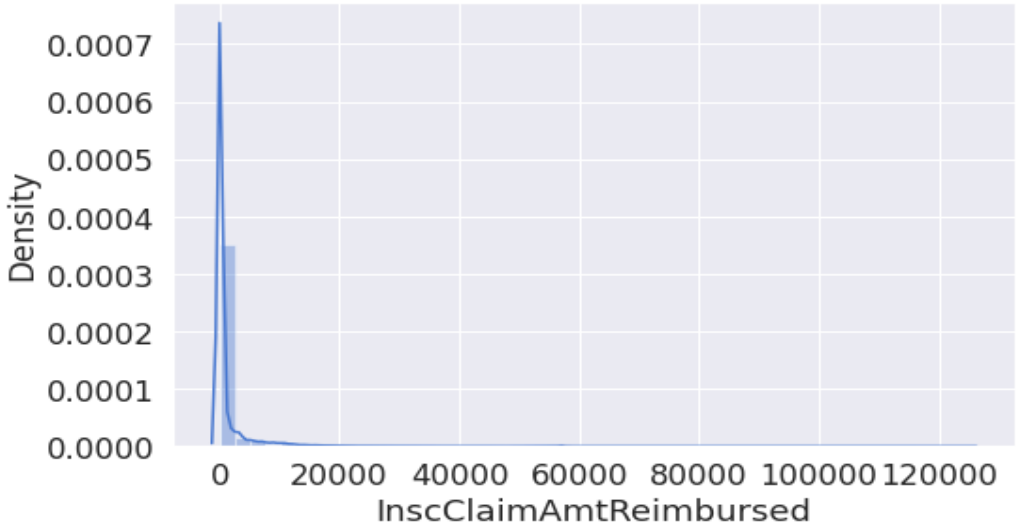

Figure 80: Money Lost in Fraud by Fraudulent Providers – Merge Data

5.6 Data Cleaning

The final step before proceeding to modelling is the selection of a workable dataset for the analysis. This consists of two tasks, minimize measurement uncertainties by cleaning out the dataset and second select only the relevant data of those providers to be analyzed. Table 4 shows the variables that discarded along with the reason why are not be chosen to conclude them in the final dataset.

Feature	Comment
BeneID	Replaced with BeneIDcount. Not important for the analysis
ClaimID	Replaced with ClaimIDcount. Not important for the analysis
AdmissionDt	Not important for the analysis
DischargeDt	Not important for the analysis
ClaimEndDt	Not important for the analysis
ClaimStartDt	Not important for the analysis
AttendingPhysician	Replaced with TotalPhysiciancount. Not important for the analysis
OperatingPhysician	Replaced with TotalPhysiciancount. Not important for the analysis
OtherPhysician	Replaced with TotalPhysiciancount. Not important for the analysis
ClmAdmitDiagnosisCode	Replaced with IsClaimCode. Not important for the analysis
ClmDiagnosisCode_1	Replaced with IsClaimCode. Not important for the analysis
ClmDiagnosisCode_2	Replaced with IsClaimCode. Not important for the analysis
ClmDiagnosisCode_3	Replaced with IsClaimCode. Not important for the analysis
ClmDiagnosisCode_4	Replaced with IsClaimCode. Not important for the analysis
ClmDiagnosisCode_5	Replaced with IsClaimCode. Not important for the analysis
ClmDiagnosisCode_6	Replaced with IsClaimCode. Not important for the analysis
ClmDiagnosisCode_7	Replaced with IsClaimCode. Not important for the analysis
ClmDiagnosisCode_8	Replaced with IsClaimCode. Not important for the analysis
ClmDiagnosisCode_9	Replaced with IsClaimCode. Not important for the analysis
ClmDiagnosisCode_10	Replaced with IsClaimCode. Not important for the analysis
DiagnosisGroupCode	Replaced with IsGroupCode. Not important for the analysis
ClmProcedureCode_5	It is not correlated with target feature or statistical significant
ClmProcedureCode_6	Big percetage of missing values. Not important for the analysis
DOB	Not important for the analysis
DOD	Not important for the analysis
Age	It is not correlated with target feature or statistical significant
Gender	It is not correlated with target feature or statistical significant
IsDead	It is not correlated with target feature or statistical significant
NoOfMonths_PartACov	It is not correlated with target feature or statistical significant
NoOfMonths_PartBCov	It is not correlated with target feature or statistical significant
ExtraDaysClaim	It is not correlated with target feature or statistical significant

Table 4: Features are Discarded from Final Dataset

5.7 Final Data

Although the initial set of behavioral metrics was quite large, it was refined to 39 that were feasible and could be applied for implementation within our research case constraints. Table 5 shows the variables that conduct the final dataset along with the reasons why we chose to choose them.

Feature	Correlation with target	Statistical Significant	Comment
PotentialFraud	1.000000	Yes	It is the target feature
Provider	-	Yes	It is statistical significant. It is necessary for predictions
BeneIDcount	0.393531	Yes	It is correlated with target feature and statistical significant
ClaimIDcount	0.374197	No	It is correlated with target feature, even it is not statistical significant
County	0.011551	Yes	It is correlated with target feature and statistical significant
Race	0.024486	Yes	It is correlated with target feature and statistical significant
State	-0.041872	Yes	It is correlated with target feature and statistical significant
DiseasesCount	0.014685	No	It is correlated with target feature, even it is not statistical significant
ChronicCond_Heartfailure	0.384131	Yes	It is correlated with target feature and statistical significant
ChronicCond_stroke	0.399206	Yes	It is correlated with target feature and statistical significant
RenalDiseaseIndicator	0.391002	Yes	It is correlated with target feature and statistical significant
ChronicCond_Alzheimer	0.380344	Yes	It is correlated with target feature and statistical significant
ChronicCond_Cancer	0.376945	Yes	It is correlated with target feature and statistical significant
ChronicCond_Diabetes	0.378881	Yes	It is correlated with target feature and statistical significant
ChronicCond_IschemicHeart	0.380093	Yes	It is correlated with target feature and statistical significant
ChronicCond_KidneyDisease	0.394239	Yes	It is correlated with target feature and statistical significant
ChronicCond_ObstrPulmonary	0.396191	Yes	It is correlated with target feature and statistical significant
ChronicCond_rheumatoidarthritis	0.380161	Yes	It is correlated with target feature and statistical significant
ChronicCond_Depression	0.377411	No	It is correlated with target feature, even it is not statistical significant
ChronicCond_Osteoporasis	0.001181	No	It is not correlated with target feature or statistical significant, but we keep all other features of procedure codes
ClmProcedureCode_1	0.076418	Yes	It is correlated with target feature and statistical significant
ClmProcedureCode_2	0.038117	Yes	It is correlated with target feature and statistical significant
ClmProcedureCode_3	0.015423	Yes	It is correlated with target feature and statistical significant
ClmProcedureCode_4	0.006820	Yes	It is statistical significant, even it is not correlated with target feature
IsInpatient	0.113401	Yes	It is correlated with target feature and statistical significant
IsClaimCode	0.053984	Yes	It is correlated with target feature and statistical significant
IsDeductible	0.088046	Yes	It is correlated with target feature and statistical significant
IsGroupCode	0.113401	Yes	It is correlated with target feature and statistical significant
DaysAdmitted	0.082004	No	It is correlated with target feature, even it is not statistical significant
DaysClaim	0.028640	No	It is correlated with target feature, even it is not statistical significant
TotalClaimCodes	0.189909	No	It is highly correlated with target feature, even it is not statistical significant
TotalPhysicians	0.037047	No	It is correlated with target feature, even it is not statistical significant
TotalProcedureCodes	0.188194	No	It is highly correlated with target feature, even it is not statistical significant
DeductibleAmtPaid	0.112016	No	It is highly correlated with target feature even it is not statistical significant
InscClaimAmtReimbursed	0.080613	No	It is highly correlated with target feature, even it is not statistical significant
IPAnnualDeductibleAmt	0.036514	No	It is correlated with target feature, even it is not statistical significant
IPAnnualReimbursementAmt	0.035027	No	It is correlated with target feature, even it is not statistical significant
OPAnnualDeductibleAmt	0.002919	No	It is not highly correlated with target feature or statistical significant, but we keep all other features of amounts
OPAnnualReimbursementAmt	0.002077	No	It is not highly correlated with target feature or statistical significant, but we keep all other features of amounts

Table 5: Features Consist of Final Dataset

5.8 Modelling and Evaluation Metrics

The actions were taken for modeling the final dataset are listed below.

- Train Test Split: For each experiment, we split the entire dataset into 80% training set and 20% test set. First, we used the training set for modeling the data, and then we used the test set to test the performance of the trained model.

- Standardizing Data using StandardScaler: Standardizing the features refers to rescaling the features so that they will have a mean of 0 and standard deviation of 1. It is a common requirement for many machine learnings models that the features should be standardized before applying the machine learning techniques. If standardization is not performed, then might affect the performance of the model. We performed standardization on all the numerical features using StandardScalar in the scikit-learn library. Standardization can be achieved as follows.

$$z = \frac{x - \mu}{\sigma}$$

$$Mean(\mu) = \frac{1}{N} \sum_{i=1}^{N} (x_i)$$

$$Standard\ deviation(\sigma) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2}$$

- Data Balancing using Synthetic Minority Over-sampling Technique (SMOTE): As mentioned earlier, the dataset is highly imbalanced. The number of legitimate transactions outnumbers the number of fraudulent transactions. In this case, if we use this dataset to train our model, the model tends to be biased towards the non-fraudulent providers, and hence, it results in the poor performance of the model when tested in unseen data. To tackle this problem, we have used the resampling technique SMOTE on the training data to make it balanced. SMOTE is a popular technique used to rebalance the dataset (Chawla, N., Bowyer, K., Hall, L., & Kegelmeyer, 2002). It aims to create new minority class examples (synthetic instances) by interpolating between several nearest minority examples rather than by oversampling with replacement, which is shown in Figure 81. As a result, it diminishes the problem of overfitting the training data. Depending upon the amount of oversampling required, nearest neighbors of minority examples are randomly selected.

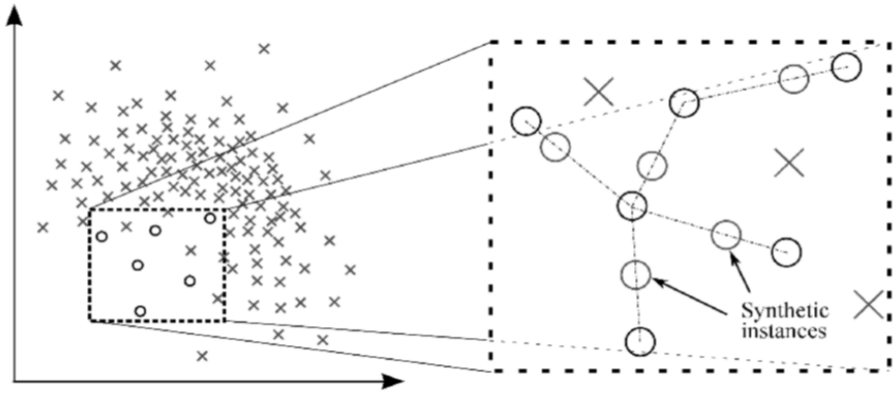

Figure 81: Generation of synthetic examples using SMOTE

- Random Forest for features reduction: Figure 82 shows the important features that were used using Random Forest.

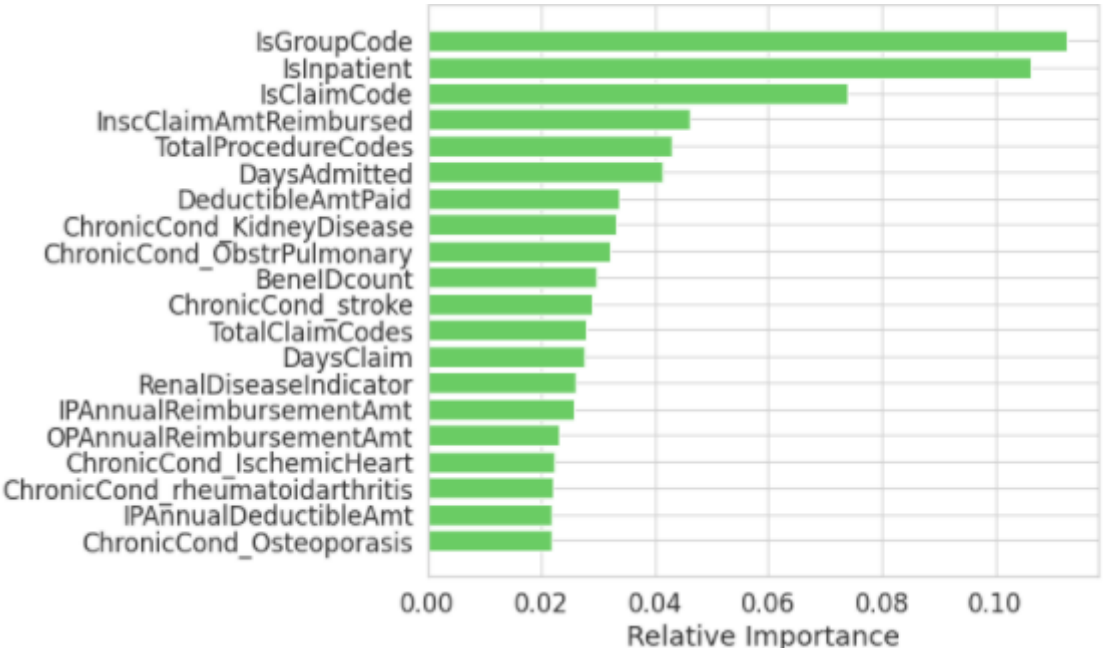

Figure 82: Important Features based on Random Forest

- Methods used: Different models are trained in order to find the one which matches the best to data. The methods that were used are:

1. Logistic Regression
2. Random Forest
3. Decision Tree Classifier
4. Support Vector Machine

As the dataset is highly imbalanced, accuracy will not be the proper metric. An important initial step will be to plot the confusion matrix. Then we need to check the misclassification which involves TP (True Positive), FP (False Positive), TN (True Negative), and FN (False Negative). FN means the cases predicted by the model are legitimate but actually it is fraudulent. FP means the cases detected by the model are fraudulent, but actually, it is legitimate. So, the evaluation metrics that were used are:

- Confusion Matrix: It is the table where TP, FP, TN, and FN counts will be plotted. From this table, we can visualize the performance of the model.
- Recall: Recall measures the ability of a classifier to determine the rate of positively marked instances that are in fact positive; therefore, for this dataset, Recall is showing the proportion that a given provider is labeled correctly.

$$Recall = \frac{TP}{TP + FN}$$

- Precision: Precision indicates how well a classifier has predicted a class by finding the ratio of actually positive instances from the pool of instances that is has marked as part of the positive class; therefore, Precision here is showing the proportion that a given provider is marked correctly against the number of providers, from the other class, marked also as the class in question.

$$Precision = \frac{TP}{TP + FP}$$

- F1 score: The F1 score is the harmonic mean or a weighted average, of both Precision and Recall and is used to organize the model performance results into one concise metric for performance comparisons, generating a number between 0 and 1, where values closer to one indicate better performance. F1 score is reasonably robust to imbalance data, it will be a correct metric for this problem.

$$F1\ score = \frac{2\ (Precision\ *\ Recall)}{(Precision\ +\ Recall)}$$

- ROC curve: Another way to examine the performance of classifiers is to use a ROC curve. A ROC graph is a curve that depicts the performance and performance tradeoff of a classification model with the False Positives along the X-axis and the True Positives along the Y-axis. The point (0, 1) is the perfect classifier: it classifies all positive cases and negative cases correctly. It is (0, 1) because the false positive FP is 0, and the TP rate is 1. The point (0, 0) represents a classifier that predicts all cases to be negative, while the point (1, 1) corresponds to a classifier that predicts every case to be positive. Point (1, 0) is the classifier that is incorrect for all classifications. A ROC curve or point is independent of class distribution or error costs. It sums up all information contained in the confusion matrix since FN is the complement of TP and TN is the complement of FP. It provides a visual tool for examining the exchange between a classifier to correctly identify positive cases and the number of negative cases incorrectly classified.

- AUC score: AUC depends on the ranking of the predicted probability score, not on absolute values. That is, AUC measures the entire two-dimensional area underneath the entire ROC curve from (0,0) to (1,1). AUC provides an aggregate measure of performance across all possible classification thresholds. One way of interpreting AUC is as the probability that the model ranks a random positive example more highly than a random negative example.

- Cost-sensitive learning: In an imbalanced classification problem like this one, the missing of a positive or minority class case is worse than incorrectly classifying an example from the negative or majority class. We can define the total cost of a classifier using this framework as the cost-weighted sum of the False Negatives (FNR) and False Positives (FPR). So, as the cost of misclassification is very high, we need to check the FPR and FNR, it should be as low as possible.

Further, the cost might be a complex multi-dimensional function, including monetary costs, reputation costs, and more. In an insurance claim example, the costs for a false positive might be the monetary cost of follow-up with the customer to the company and the cost of a false negative might be the cost of the insurance claim, and we might assign no cost to correct predictions in each class. The purpose of cost-sensitive learning is to build a model with minimum misclassification costs, which is the total cost. The below value is what we seek to minimize in cost-sensitive learning.

$$Total\ Cost\ =\ Cost\ of\ FN\ *\ False\ Negatives\ +\ Cost\ of\ FP\ *\ False\ Positives$$

5.9 Models Comparison

Below are the four different approaches that were followed in this case study and Tables 6, 7, 8, and 9 summarize the results.

- Approach 1: Original Data with all features
- Approach 2: Original Data with important features based on Random Forest
- Approach 3: Balancing Data using SMOTE with all features
- Approach 4: Balancing Data with important features based on Random Forest.

From Table 6 we can observe that performance of Decision Tree Classifier is not good. Logistic Regression, Random Forest, and SVM seem to have a good fit on the data, but Logistic Regression superiors in relation to other models based on performances.

Metrics	LR Train	LR Test	RF Train	RF Test	DTC Train	DTC Test	SVM Train	SVM Test
F1score	0.563877	0.612795	0.546012	0.562500	0.598597	0.524096	0.513644	0.629630
Recall	0.790123	0.900990	0.879012	0.891089	0.948148	0.861386	0.395062	0.504950
Precision	0.438356	0.464286	0.395996	0.410959	0.437358	0.376623	0.733945	0.836066
AUC	0.932976	0.958226	0.947552	0.953583	0.963882	0.906062	0.932255	0.955683

Table 6: Approach 1 - All Features – No Sampling

From Table 7 we can see that using all features after balancing the data with SMOTE there is no improvement in performances for all models.

Metrics	LR Train	LR Test	RF Train	RF Test	DTC Train	DTC Test	SVM Train	SVM Test
F1score	0.736499	0.619718	0.735527	0.549383	0.923149	0.513208	0.670303	0.642487
Recall	0.807143	0.871287	0.901020	0.881188	0.99898	0.673267	0.564286	0.613861
Precision	0.677226	0.480874	0.621393	0.399103	0.858019	0.414634	0.825373	0.673913
AUC	0.941161	0.957570	0.952381	0.952837	0.994203	0.794840	0.940574	0.955804

Table 7: Approach 2 - All Features – Sampling

Logistic Regression and SVM seem to have a good fit on the data. Random Forest has a small decrease in AUC score and precision, and the Decision Tree Classifier is overfitting the data.

Metrics	LR Train	LR Test	RF Train	RF Test	DTC Train	DTC Test	SVM Train	SVM Test
F1score	0.564872	0.596610	0.548114	0.547692	0.530314	0.51632	0.512903	0.629630
Recall	0.790123	0.871287	0.879012	0.881188	0.896296	0.861386	0.392593	0.504950
Precision	0.43956	0.453608	0.398210	0.397321	0.376556	0.368644	0.739535	0.836066
AUC	0.932224	0.958731	0.946541	0.952837	0.937853	0.936284	0.931299	0.957560

Table 8: Approach 3 - Important Features – No Sampling

From Table 9 we can see that using only important features based on Random Forest after balancing the data with SMOTE there is no improvement in performances for all models.

Metrics	LR Train	LR Test	RF Train	RF Test	DTC Train	DTC Test	SVM Train	SVM Test
F1score	0.742086	0.619718	0.737589	0.539394	0.985420	0.491667	0.667071	0.649485
Recall	0.813265	0.871287	0.902041	0.881188	1.000000	0.584158	0.561224	0.623762
Precision	0.682363	0.480874	0.623853	0.388646	0.971259	0.424460	0.822123	0.677419
AUC	0.942016	0.958317	0.951921	0.952493	0.999891	0.755453	0.941407	0.957540

Table 9: Approach 4 - Important Features – Sampling

All in all, approach 1 has better performance than approach 2. So, it can be said that the model which is performing better is Logistic Regression using non-balancing data with all features.

5.10 Final Model and Interpretation of Results

After the evaluation, it is found that Logistic Regression without using SMOTE method with all features performed the best in test data with an F1 score of 0.612795, recall score of 0.900990, a precision score of 0.464286 and AUC score of 0.958226. Below is presented the confusion matrix along with the evaluation metrics for the chosen approach for training and test data.

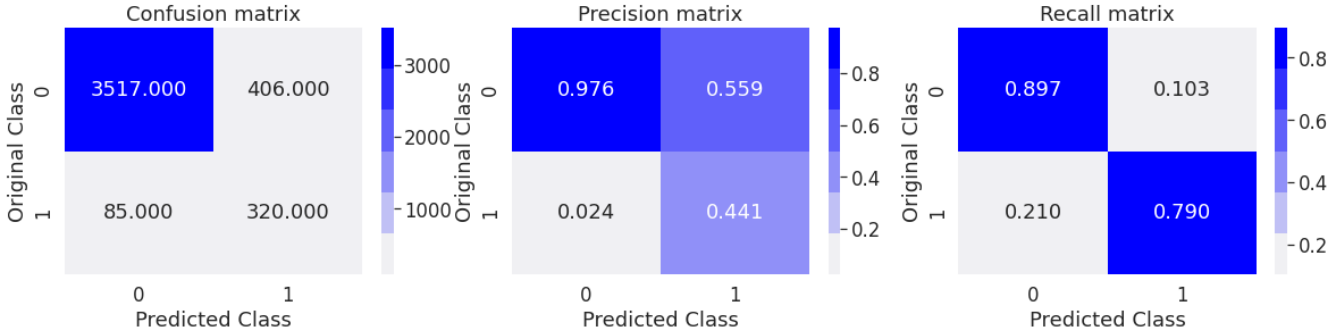

Figure 83: Train Metrics of Final Model

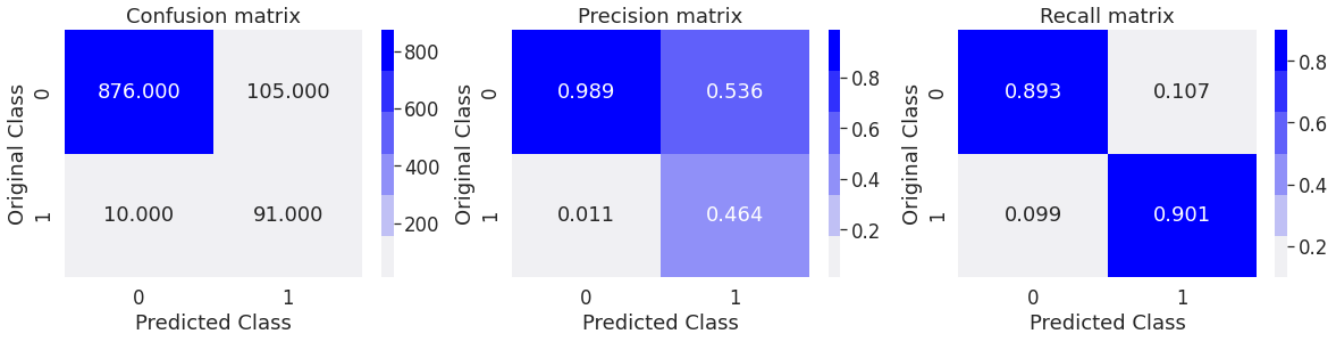

Figure 84: Test Metrics of Final Model

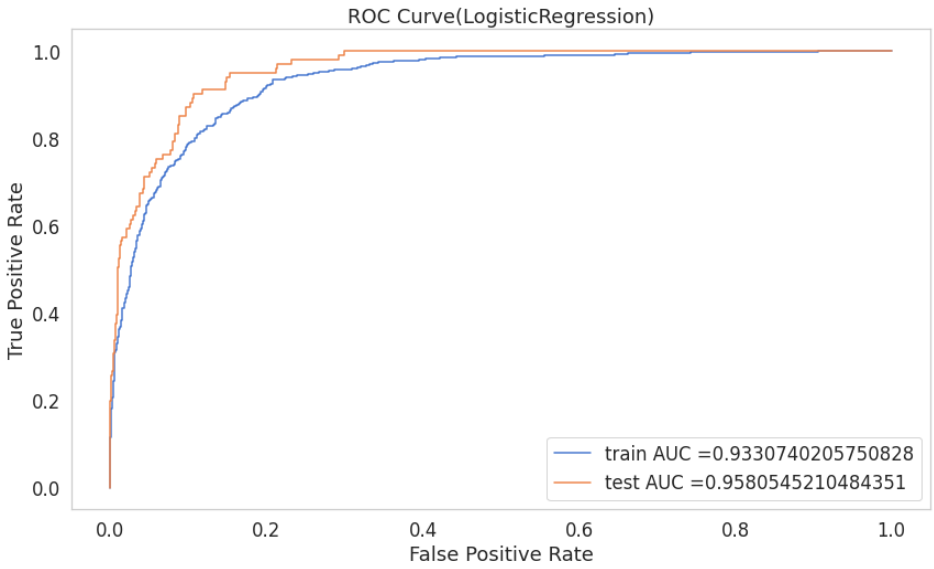

Figure 85: Train and Test Roc Curve of Final Model

Applying the above model, which has an AUC score of 0.96, means that the correct/missing classification prediction is 96%, or in other words, every time this model predicts that a provider is fraudulent there is a 96% odd that this provider is indeed fraudulent.

Furthermore, our final model has a recall score of 0.90, which means that the True Positive Rate is 90%. In simple terms, 91 fraudulent providers were found out of the 105 true positive cases. The sum of reimbursement amount for the 91 providers that were correctly detected as fraudulent is 71,014,410, which is around 71 million out of the total 556 million.

Also, if we assume a cost of $5 for FP and a cost of $88 for FN the total cost of the training dataset is calculated as below:

*Total Cost of Train Dataset = 88 * 406 + 5 * 85=36,153.*

The total cost of the test dataset is calculated as below:

*Total Cost of Test Dataset = 88 * 105 + 5 * 10=9,290.*

This means that the final model accomplished a reduction of 25.70% in the total cost of fraud.

5.11 Final Pipeline

As was referred to before Logistic Regression without using SMOTE method with all features worked the best for this healthcare provider's fraud detection problem. In the Appendix is the code snippet of the final pipeline. The concerned reader can contact me to ask me for the complete pipeline.

5.12 Chapter Summary

In the last chapter, we applied machine learning techniques to predict whether a health insurance claim is fraudulent or not. For this, we collected a publicly available dataset provided by the machine learning group of Kaggle, which contains the transactions of inpatient claims, outpatient claims, and beneficiary details. It contains 5410 transactions out of which only 506 are fraudulent.

The dataset is highly imbalanced as the positive class accounts for only 9.35% of the total transactions. When providing input data of a highly imbalanced class distribution to the predictive model, the model tends to be biased towards the majority of samples. As a result, it tends to misrepresent a fraudulent transaction as a genuine transaction. To tackle this problem, we implemented a data level approach which includes the resampling technique SMOTE. In addition, based on Random Forest it was found the 20 most important features for this dataset.

In addition, the algorithms selected to perform this fraud case are Logistic Regression, Random Forest, Decision Tree Classifier, and Support Vector Machine. Then, we train all four algorithms with and without using the resampling technique SMOTE, as also using all features and using only important features based on random forest. The comparison of the results revealed that Logistic Regression using all features without resampling the data approach performed better than other models. This approach manages to detect correctly 91 out of 105 fraudulent providers with an f1 score of 0.61, AUC score of 0.96 and decrease the total cost by 25.70%.

All in all, the disastrous effect of health care fraud needs to be reduced. The data analytics techniques may not completely eliminate fraud but surely reduced it. With the help of machine learning techniques, the prevention of huge losses can be incurred by health insurance companies or the governments by identifying health care providers who make fraud claims on behalf of their beneficiaries and helping them to disburse insurance money to beneficiaries who truly deserve it. Finally, this will also help in bringing down the premium costs for insurance and thereby making healthcare more affordable.

APPENDIX

In [1]:

```
import numpy as np
import pandas as pd
import os
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.metrics import f1_score, roc_curve, precision_score, recall_score, confusion_
matrix, auc
from sklearn.metrics import confusion_matrix,accuracy_score,cohen_kappa_score,roc_auc_scor
e,f1_score,auc
from sklearn.model_selection import GridSearchCV
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegressionCV
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
import pprint
import scipy as sc
import pandas_profiling as profile
import warnings
warnings.filterwarnings("ignore")
from sklearn.preprocessing import StandardScaler,MinMaxScaler
import pickle
from scipy import stats
import tensorflow as tf
from pylab import rcParams
from keras.models import Model, load_model
from keras.layers import Input, Dense
from keras.callbacks import ModelCheckpoint, TensorBoard
from keras import regularizers
from sklearn.linear_model import LogisticRegression
from sklearn import metrics
from sklearn.metrics import roc_curve, auc , f1_score
from sklearn.metrics import confusion_matrix
%matplotlib inline
sns.set(style='whitegrid', palette='muted', font_scale=1.5)
rcParams['figure.figsize'] = 14, 8
RANDOM_SEED = 42
pd.set_option("display.max_rows", 5, "display.max_columns", None)
import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))
```

```
/kaggle/input/healthcare-provider-fraud-detection-analysis/Test-1542969243754.csv
/kaggle/input/healthcare-provider-fraud-detection-analysis/Train_Beneficiarydata-15428
65627584.csv
/kaggle/input/healthcare-provider-fraud-detection-analysis/Train_Inpatientdata-1542865
627584.csv
/kaggle/input/healthcare-provider-fraud-detection-analysis/Test_Outpatientdata-1542969
243754.csv
/kaggle/input/healthcare-provider-fraud-detection-analysis/Train-1542865627584.csv
/kaggle/input/healthcare-provider-fraud-detection-analysis/Test_Beneficiarydata-154296
9243754.csv
/kaggle/input/healthcare-provider-fraud-detection-analysis/Test_Inpatientdata-15429692
43754.csv
/kaggle/input/healthcare-provider-fraud-detection-analysis/Train_Outpatientdata-154286
5627584.csv
```

In [2]:

```
train=pd.read_csv('../input/healthcare-provider-fraud-detection-analysis/Train-1542865627584.csv')
test=pd.read_csv('../input/healthcare-provider-fraud-detection-analysis/Test-1542969243754.csv')
train_inpatient=pd.read_csv('../input/healthcare-provider-fraud-detection-analysis/Train_Inpatientdata-1542865627584.csv')
train_outpatient=pd.read_csv('../input/healthcare-provider-fraud-detection-analysis/Train_Outpatientdata-1542865627584.csv')
test_inpatient=pd.read_csv('../input/healthcare-provider-fraud-detection-analysis/Test_Inpatientdata-1542969243754.csv')
test_outpatient=pd.read_csv('../input/healthcare-provider-fraud-detection-analysis/Test_Outpatientdata-1542969243754.csv')
train_beneficiary=pd.read_csv('../input/healthcare-provider-fraud-detection-analysis/Train_Beneficiarydata-1542865627584.csv')
test_beneficiary=pd.read_csv('../input/healthcare-provider-fraud-detection-analysis/Test_Beneficiarydata-1542969243754.csv')
```

In [3]:

```
def combine_data(provider, beneficiary, inpatient, outpatient):
    ''' This function combines provider, beneficiary, inpatient and outpatient data
        together in a single file. '''
    # Finding common columns in inpatient data and outpatient data
    common_inp_outp = [col for col in inpatient if col in outpatient]
    # Outer joining inpatient data and outpatient data using common columns
    train_in_out = outpatient.merge(inpatient, on=common_inp_outp, how='outer', indicator='IsInpatient')
    train_in_out = train_in_out.replace({'left_only':0, 'right_only':1})
    # Inner joining beneficary data and combination of inpatient and outpatient
    train_benef_in_out = beneficiary.merge(train_in_out, on='BeneID')
    # Combining all combinations with train data containing target variable
    combined_data = provider.merge(train_benef_in_out, on='Provider')
    return combined_data
```

In [4]:

```
# Combine all data from training set
train_data = combine_data(train, train_beneficiary, train_inpatient, train_outpatient)
```

In [5]:

```
# 1. Replacing Yes with 1 and No with 0 in PotentialFraud column
train_data['PotentialFraud'] = train_data['PotentialFraud'].replace({'Yes':1, 'No':0})

# 2. Changing DOB and DOD to date type
train_data['DOB'] = pd.to_datetime(train_data['DOB'])
train_data['DOD'] = pd.to_datetime(train_data['DOD'])

# 3. Replacing 1 with Female and 2 with Male
train_data['Gender'] = train_data['Gender'].replace({1:'Female', 2:'Male'})

# 4. Replacing '0' with 0 and 'Y' with 1 in RenalDiseaseIndicator column
train_data.RenalDiseaseIndicator.replace({'0':0, 'Y':1}, inplace=True)

# 5. Replacing 2 with 0 for all chronic conditions for eg: ChronicCond_Alzheimer
train_data.replace({'ChronicCond_Alzheimer': 2, 'ChronicCond_Heartfailure': 2, 'ChronicCond_KidneyDisease': 2,
                    'ChronicCond_Cancer': 2, 'ChronicCond_ObstrPulmonary': 2, 'ChronicCond_Depression': 2,
                    'ChronicCond_Diabetes': 2, 'ChronicCond_IschemicHeart': 2, 'ChronicCond_Osteoporasis': 2,
                    'ChronicCond_rheumatoidarthritis': 2, 'ChronicCond_stroke': 2 }, 0, inplace=True)

# 6. Changing data type of ClaimStartDt and ClainEndDt to datetime
train_data['ClaimStartDt'] = pd.to_datetime(train_data['ClaimStartDt'])
train_data['ClaimEndDt'] = pd.to_datetime(train_data['ClaimEndDt'])
```

```
# 7. Changing data type of AdmissionDt and DischargeDt to datetime
train_data['AdmissionDt'] = pd.to_datetime(train_data['AdmissionDt'])
train_data['DischargeDt'] = pd.to_datetime(train_data['DischargeDt'])

# 8. Convert type of State and Race to categorical
train_data.Gender=train_data.State.astype('category')
train_data.Race=train_data.Race.astype('category')

# 1. Age
dod_year = train_data['DOD'].dt.year
start_year = train_data['ClaimStartDt'].dt.year
end_year = train_data['ClaimEndDt'].dt.year
train_data['Age'] = 2009 - train_data['DOB'].dt.year

# 2. IsDead
train_data['IsDead'] = train_data['DOD'].apply(lambda x: 0 if pd.isnull(x) else 1)

# 3. DaysAdmitted

train_data['DaysAdmitted'] = (train_data['DischargeDt'] - train_data['AdmissionDt']).filln
a(pd.Timedelta('0 days'))
train_data['DaysAdmitted'] = train_data['DaysAdmitted'].apply(lambda x: int(str(x).split()
[0]))
train_data['DaysAdmitted'][train_data.IsInpatient == 1]

# 4. DiseasesCount
cols = ['ChronicCond_Alzheimer','ChronicCond_Heartfailure',
        'ChronicCond_KidneyDisease','ChronicCond_Cancer',
        'ChronicCond_ObstrPulmonary','ChronicCond_Depression',
        'ChronicCond_Diabetes','ChronicCond_IschemicHeart',
        'ChronicCond_Osteoporasis','ChronicCond_rheumatoidarthritis','ChronicCond_stroke']

train_data['DiseasesCount'] = train_data[cols].sum(axis=1)

# 5. TotalPhysicians
train_data['TotalPhysicians'] = train_data[['AttendingPhysician', 'OperatingPhysician', 'O
therPhysician']]\
                                          .apply(lambda x: sum(pd.notnull(x)), axis=1)

# 6. TotalClaimCodes
cols = ['ClmDiagnosisCode_1', 'ClmDiagnosisCode_2', 'ClmDiagnosisCode_3',
       'ClmDiagnosisCode_4', 'ClmDiagnosisCode_5', 'ClmDiagnosisCode_6',
       'ClmDiagnosisCode_7', 'ClmDiagnosisCode_8', 'ClmDiagnosisCode_9','ClmDiagnosisCode_
10']
train_data['TotalClaimCodes'] = train_data[cols].apply(lambda x: sum(pd.notnull(x)), axis=
1)

# 7. TotalProcedureCodes
cols = ['ClmProcedureCode_1', 'ClmProcedureCode_2','ClmProcedureCode_3',
        'ClmProcedureCode_4', 'ClmProcedureCode_5','ClmProcedureCode_6']
train_data['TotalProcedureCodes'] = train_data[cols].apply(lambda x: sum(pd.notnull(x)), a
xis=1)

# 8. DaysClaim
train_data['DaysClaim'] = (train_data['ClaimEndDt'] - train_data['ClaimStartDt']).fillna(p
d.Timedelta('0 days'))
train_data['DaysClaim'] = train_data['DaysClaim'].apply(lambda x: int(str(x).split()[0]))

# 9. ExtraDaysClaim
(train_data.DaysClaim > train_data.DaysAdmitted).value_counts()
train_data['ExtraDaysClaim'] = (train_data.DaysClaim > train_data.DaysAdmitted)\
                                              .replace({True:1, False:0})
```

```
# 10. IsClaimCode
train_data['IsClaimCode'] = train_data['ClmAdmitDiagnosisCode'].apply(lambda x: 0 if pd.is
null(x) else 1)

# 11. IsGroupCode
train_data['IsGroupCode'] = train_data['DiagnosisGroupCode'].apply(lambda x: 0 if pd.isnul
l(x) else 1)

# 12. IsDeductible
train_data['IsDeductible'] = train_data['DeductibleAmtPaid'].apply(lambda x: 0 if pd.isnul
l(x) else 1)

# Replacing null values with 0 in DeductibleAmtPaid
train_data['DeductibleAmtPaid'] = train_data['DeductibleAmtPaid'].fillna(0)

# Drop
remove_columns=['DOB','DOD','Age','Gender','IsDead','NoOfMonths_PartACov','NoOfMonths_Part
BCov','AdmissionDt','DischargeDt','AttendingPhysician','ClaimEndDt','ClaimStartDt','ClmAdm
itDiagnosisCode','ClmDiagnosisCode_1','ClmDiagnosisCode_10','ClmDiagnosisCode_2','ClmDiagn
osisCode_3','ClmDiagnosisCode_4','ClmDiagnosisCode_5','ClmDiagnosisCode_6','ClmDiagnosisCo
de_7','ClmDiagnosisCode_8','ClmDiagnosisCode_9','DiagnosisGroupCode','OperatingPhysician',
'OtherPhysician','ClmProcedureCode_5','ClmProcedureCode_6','ExtraDaysClaim']
train_cleaned=train_data.drop(columns=remove_columns, axis=1)

# Do one hot encoding for State and Race
train_cleaned=pd.get_dummies(train_cleaned,columns=['State','Race'])

# Storing the target variable of each Provider Id
output_train = train_cleaned[['Provider', 'PotentialFraud']].drop_duplicates(subset='Provi
der')

# Grouping columns with aggregation=count on Provider column
bene_count = train_cleaned.groupby('Provider').BeneID.nunique().reset_index().rename(colum
ns={'BeneID':'BeneIDcount'})
claim_count = train_cleaned.groupby('Provider').ClaimID.nunique().reset_index().rename(col
umns={'ClaimID':'ClaimIDcount'})
agg_count_data = bene_count.merge(claim_count, on='Provider', how='inner')

# Grouping columns with aggregation=sum on Provider column
agg_sum_data = train_cleaned.groupby('Provider')[['RenalDiseaseIndicator',
       'ChronicCond_Alzheimer', 'ChronicCond_Heartfailure',
       'ChronicCond_KidneyDisease', 'ChronicCond_Cancer',
       'ChronicCond_ObstrPulmonary', 'ChronicCond_Depression',
       'ChronicCond_Diabetes', 'ChronicCond_IschemicHeart',
       'ChronicCond_Osteoporasis', 'ChronicCond_rheumatoidarthritis',
       'ChronicCond_stroke', 'IsInpatient', 'IsClaimCode',
        'IsGroupCode', 'IsDeductible']].sum().reset_index()

# Grouping columns with aggregate=mean on Provider column
agg_mean_data = train_cleaned.groupby('Provider')[['IPAnnualReimbursementAmt',
       'IPAnnualDeductibleAmt', 'OPAnnualReimbursementAmt',
       'OPAnnualDeductibleAmt', 'InscClaimAmtReimbursed',
       'DeductibleAmtPaid', 'DaysAdmitted',
       'DiseasesCount', 'TotalPhysicians', 'TotalClaimCodes',
       'TotalProcedureCodes', 'DaysClaim']].mean().reset_index()

# Merging all aggregated groups and target column
train_grouped = agg_count_data.merge(agg_sum_data, on='Provider', how='inner')\
                        .merge(agg_mean_data, on='Provider', how='inner')\
                        .merge(output_train, on='Provider', how='inner')
```

In [6]:

```
# Confusion matrix
def plot_confusion_matrix(test_y, predict_y):
```

```
    C = confusion_matrix(test_y, predict_y)
    A =(((C.T)/(C.sum(axis=1))).T)
    B =(C/C.sum(axis=0))
    plt.figure(figsize=(20,4))

    labels = [0,1]
    # representing A in heatmap format
    cmap=sns.light_palette("blue")
    plt.subplot(1, 3, 1)
    sns.heatmap(C, annot=True, cmap=cmap, fmt=".3f", xticklabels=labels, yticklabels=label
s)
    plt.xlabel('Predicted Class')
    plt.ylabel('Original Class')
    plt.title("Confusion matrix")

    plt.subplot(1, 3, 2)
    sns.heatmap(B, annot=True, cmap=cmap, fmt=".3f", xticklabels=labels, yticklabels=label
s)
    plt.xlabel('Predicted Class')
    plt.ylabel('Original Class')
    plt.title("Precision matrix")

    plt.subplot(1, 3, 3)
    # representing B in heatmap format
    sns.heatmap(A, annot=True, cmap=cmap, fmt=".3f", xticklabels=labels, yticklabels=label
s)
    plt.xlabel('Predicted Class')
    plt.ylabel('Original Class')
    plt.title("Recall matrix")

    plt.show()

# Validation
def validate_model(clf, X_train, X_test, y_train, y_test):
    # predict the probability of train data
    y_train_pred = pred_prob(clf, X_train)
    # predict the probability of test data
    y_test_pred = pred_prob(clf, X_test)
    # calculate tpr, fpr for diffeent thresholds using roc_curve
    train_fpr, train_tpr, tr_thresholds = roc_curve(y_train, y_train_pred)
    test_fpr, test_tpr, te_thresholds = roc_curve(y_test, y_test_pred)

    # calculate auc for train and test
    train_auc = auc(train_fpr, train_tpr)
    print("Train AUC = ", train_auc)
    test_auc = auc(test_fpr, test_tpr)
    print("Test AUC = ", test_auc)

    draw_roc(train_fpr, train_tpr, test_fpr, test_tpr)

    best_t = find_best_threshold(tr_thresholds, train_fpr, train_tpr)

    train_prediction, test_prediction = draw_confusion_matrix(best_t, X_train, X_test, y_t
rain, y_test, y_train_pred, y_test_pred)


    train_f1_score = f1_score(y_train, train_prediction)
    test_f1_score = f1_score(y_test, test_prediction)

    return test_auc, test_f1_score, best_t
```

In [7]:

```
# Split data in train and validation
df = train_grouped
```

```
X =  df.drop(['Provider', 'PotentialFraud'], axis=1)
Y = df['PotentialFraud'].values
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.2, stratify=Y, rando
m_state=42)
df1, df2 = train_test_split(df, test_size=0.2, stratify=Y, random_state=42)

# Standardizing data
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Modelling
c_range = np.logspace(-4, 4, 20)

clf = GridSearchCV(LogisticRegression(class_weight='balanced', max_iter=1000),
                   param_grid = {'C':c_range},
                   scoring='f1',
                   n_jobs=-1)

clf.fit(X_train_scaled, y_train)
best_log_model = clf.best_estimator_
```

In [17]:

```
# Predictions
df2['Prediction']=y_test_pred
df2.loc[:,["Provider","PotentialFraud","Prediction"]]
```

Out[17]:

	Provider	PotentialFraud	Prediction
1266	PRV52573	0	0
1394	PRV52731	0	0
...	...	...	...
4275	PRV56346	0	0
4060	PRV56080	0	0

1082 rows × 3 columns

BIBLIOGRAPHY

Chawla, N., Bowyer, K., Hall, L., & Kegelmeyer, (2002). *Smote: Synthetic minority over-sampling technique*, Journal of Artificial Intelligence Research.

Christopher M. Bishop, (2006). *Pattern Recognition and Machine Learning*, Springer.

Dallas Thorntona, Michel Brinkhuis, Chintan Amrit, Robin Aly (2015). *Categorizing and Describing the Types of Fraud in Healthcare, tackling fraud and reducing waste,* Conference on Enterprise Information Systems / International Conference on Project Management / Conference on Health and Social Care Information Systems and Technologies

E.W.T. Ngai, Yong Hu, Y.H. Wong, Yijun Chen, Xin Sun, (2010). *The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature*, Institute of Business Intelligence and Knowledge Discovery, Department of E-commerce, Guangdong University of Foreign Studies, Sun Yat-Sen University, Guangzhou 510006, PR China.

European Commission – Directorate-General Home Affairs (2013). *Study on Corruption in the Healthcare Sector*, HOME/2011/ISEC/PR/047-A2 October 2013

Guido Cornelis van Capelleveen, (2013). *Outlier based Predictors for Health Insurance Fraud Detection within U.S. Medicaid*, M.Sc. Thesis, University of Twente & University of California, San Diego.

H. Lookman Sithic, T. Balasubramanian (2013). *Survey of Insurance Fraud Detection Using Data Mining Techniques*, International Journal of Innovative Technology and Exploring Engineering (IJITEE).

Isaac Akomea (2019). *Causes, effects, and deterrence of insurance fraud: evidence from Ghana*, Emerald Insight

Jenny Wilson, (2019). *Major Types of Health Insurance Frauds and Their Punishments*, https://sybridmd.com/, Empowering Medical Practices.

Jim Gee, Dr. Mark Button, Graham Brooks (2015). *The financial cost of healthcare fraud 2015*, PKF (LLP) and University of Portsmouth

Jing Li, Kuei-Ying Huang, Jionghua Jin, Jianjun Shi, (2008). *A survey on statistical methods for health care fraud detection*, Springer Science + Business Media, LLC 2007.

João Amorim Queirós Galamba de Oliveira, (2017). *Detection of Fraud and Corruption in Healthcare System*, M.Sc. Thesis, Electrical and Computer Engineering, Tecnico Lisboa.

Karca Duru Aral, (2009). *Prescription fraud detection via data mining: A methodology proposal*, M.Sc. Thesis, Bilkent University.

Lexis Nexis (2011). *Bending the Cost Curve: Analytics-Driven Enterprise Fraud Control*

M.Akhil Jabbar, B.L Deekshatulua, Priti Chandra (2013). *Classification of Heart Disease Using K-Nearest Neighbor and Genetic Algorithm*, International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA) 2013, India.

Margit Sommersguter-Reichmann, Claudia Wild, Adolf Stepan, Gerhard Reichmann, Andrea Fried (2018). *Individual and Institutional Corruption in European and US Healthcare: Overview and Link of Various Corruption Typologies*, Appl Health Econ Health Policy

Matthew Andrew Herland (2019). *Big Data Analytics and Engineering for Medicare Fraud Detection*, M.Sc. Thesis, Florida Atlantic University.

Paul Vincke (2016). *Fighting Fraud & Corruption in Healthcare in Europe: a work in progress*, European Healthcare Fraud and Corruption Network, Portugal

Paul Vincke, Jonathan Cylus (2011). *Health care fraud detection and corruption Europe: An overview*, Eurohealth incorporating Euro Observer.

Rebecca S. Busch, (2008). *Healthcare Fraud Auditing and Detection Guide*, John Wiley and Sons, Inc., Hoboken, New Jersey.

Rekha Bhowmik (2008). *Data Mining Techniques in Fraud Detection*, ADFSL Conference on Digital Forensics, Security and Law, Department of Computer Science Sam Houston State University Huntsville, Texas

Rekha Bhowmik (2008). *Data Mining Techniques in Fraud Detection*, University of Texas at Dallas, USA.

Ronish Shakya (2013). *Application of Machine Learning Techniques in Credit Card Detection*, M.Sc. Thesis, University of Nevada, Las Vegas.

Sharifa Rigga Mambo, (2019). *Use of Data Mining to detect fraud health insurance claims*, M.Sc. Thesis, University of Nairobi School of Computing and Informatics.

Shivani S. Waghade, (2018). *A Comprehensive Study of Healthcare Fraud Detection based on Machine Learning*, International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 6 (2018) pp. 4175-4178, Research India Publications.

Simon Haykin, (2009). *Neural Networks and Learning Machines*, McMaster University Hamilton, Ontario, Canada.

Tahir Ekin, Francesca Ieva, Fabrizio Ruggeri, Refik Soyer (2018). *Statistical Medical Fraud Assessment: Exposition to an Emerging Field*, The Institute for Integrating Statistics in Decision Sciences.

Technical Report TR-2017-2, 1McCoy College of Business, Texas State University, Department of Mathematics, Politecnico di Milano, CNR-IMATI, School of Business, The George Washington University.

Timo Koski, John M. Noble (2009). *Bayesian Networks: An Introduction*, A John Wiley and Sons, Ltd., Publication, United Kingdom.

Vipula Rawte, G Anuradha (2015). *Fraud Detection in Health Insurance using Data Mining Techniques, 2015 International Conference on Communication*, Information and Computing Technology (ICCICT), Jan. 16-17, Mumbai, India.

Yufeng Kou, Chang-Tien Lu, Sirirat Sinvongwattana, Yo-Ping Huang (2004). *Survey of Fraud Detection Techniques*, Proceedings of the 2004 IEEE International Conference on Networking, Sensing and Control Taipei, Taiwan, March 21-23, 2004.

Internet References

IBM Cloud Education (2020), Neural Networks, https://www.ibm.com.

Abhinav Ralhan (2018), Self Organizing Maps, https://www. medium.com.

An Introduction to Cluster Analysis, https://www. alchemer.com.

A Guide to Principal Component Analysis (PCA) for Machine Learning, https://www.keboola.com.

A Step-by-Step Explanation of Principal Component Analysis (PCA), https://www. builtin.com.

What Is Principal Component Analysis (PCA) and How It Is Used?, https://www.sartorius.com.

Health Insurance Fraud Definition, Types, Impact and Ways To Reduce It, https://www.insurancesamadhan.com.

Cost-Sensitive Learning for Imbalanced Classification (2020), https://machinelearningmastery.com.

Anik Manik (2021), Healthcare Provider Fraud Detection Analysis using Machine Learning, https://medium.com.

Sumeet Shahu (2021), Healthcare Provider Fraud Detection, https://medium.com.