
Tackling Fraud, Waste and Abuse in Claims Processing

DSEOG ZG628T: Dissertation

by

Soumya Ranjan Pati

BITS ID No:2022OG04006

Dissertation work carried out at Optum, India

Submitted in partial fulfillment of
M.Tech. Data Science and Engineering degree programme

Under the Supervision of
Mohini Rastogi, Mentor, Senior Data Analyst
Optum Global Solutions, Gurgaon



**BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE
PILANI (RAJASTHAN)**

May, 2024

CERTIFICATE

This is to certify that the Dissertation entitled "Tackling Fraud, Waste and Abuse in Claims Processing" and submitted by Soumya Ranjan Pati having ID-No. 2022OG04006 for the partial fulfillment of the requirements of M.Tech. Data Science and Engineering degree of BITS-Pilani, embodies the bonafide work done by him/her under my supervision.



Place: Hyderabad

Date: 17/05/2024

Signature of the Mentor/Supervisor

Name: Mohini Rastogi

Designation: Senior Data Analyst

Organization: Optum Global Solutions

Location: Gurgaon

Birla Institute of Technology & Science, Pilani
Work-Integrated Learning Programmes Division
Fourth Semester 2022-2024
DSEOG ZG628T: Dissertation

ABSTRACT

BITS ID No. : 2022OG04006

NAME OF THE STUDENT : SOUMYA RANJAN PATI

EMAIL ADDRESS : 2022og04006@wilp.bits-pilani.ac.in

STUDENT'S EMPLOYING ORGANIZATION & LOCATION : Optum Global Solutions, Hyderabad

SUPERVISOR'S NAME : Mohini Rastogi

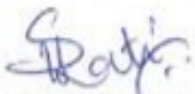
SUPERVISOR'S EMPLOYING ORGANIZATION & LOCATION : Optum Global Solutions, Gurgaon

SUPERVISOR'S EMAIL ADDRESS: Mohini.rastogi@optum.com

DISSERTATION TITLE : Tackling Fraud, Waste and Abuse in Claims Processing

ABSTRACT : The problem statement that I am trying to address as part of this dissertation is to develop a framework by leveraging data mining and machine learning concepts to tackle fraud, waste, and abuse (FWA) in claims processing. Despite existing measures, FWA continues to drain resources, inflate costs, and undermine the integrity of the healthcare systems. The goal is to identify and prevent fraudulent activities, minimize unnecessary expenses due to waste and mitigate abuse in the claims processing systems. Additionally, the company can generate extra revenue by recovering the extra dollars from the providers by re-processing the incorrectly paid claims.

Broad Domain of Work: Data Mining in health care industry



(Signature of Student)

Date: 16-05-2024



(Signature of Mentor)

Date: 17-05-2024

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my mentors Mohini Rastogi and Shashi Khan, whose guidance, insightful feedback, and unwavering support have been the cornerstone of my project work. Their expertise and encouragement were instrumental in shaping my research and empowering me to delve deeper into the complexities of fraud, waste, and abuse in claims processing.

A special thanks goes to the operations team member Sitakanta Nayak, whose collaboration was vital in the data gathering process. Being a business analyst himself, his comprehensive understanding of the business requirements, identification of gaps in the current process, and in-depth knowledge of the problem areas provided me with the essential context needed to undertake this dissertation.

The development of different models to predict potential fraudulent claims could not have been possible without their contributions. Their dedication to excellence and commitment to supporting my endeavors have left an indelible mark on my work.

I am also immensely grateful to my family, whose constant love and support have been my source of strength and motivation. Their belief in my abilities has been a constant source of encouragement throughout this journey.

Furthermore, I extend my appreciation to the resources and facilities provided by my company Optum Global Solutions and my college Birla Institute of Technology and Science, Pilani. The environment and opportunities furnished by my company and institution have been pivotal in the successful completion of my dissertation.

Table of Contents

Chapter 1. Problem Statement and Scope of Dissertation	1
Chapter 2: Details on the Progress	4
2.1. Data Acquisition and Pre-processing:.....	4
Figure 1. Tables and Columns for FraudAnalysis and WastageAnalysis	4
2.1.1 Pre-processing and Analysis on “Wastage Analysis” dataset.....	4
2.1.2 Pre-processing and Analysis on “Fraud Analysis” dataset	6
2.2 Model Development:.....	6
Chapter-3: What is remaining and plan to complete it?	13
List of abbreviations / Acronyms.....	13
Summary.....	14
Literature Surveys	15
Appendices	16
Definition of the commonly used metrics to evaluate the model performance:	16
Checklist of items for the Dissertation Report	18

Table of Figures

Figure 1. Tables and Columns for FraudAnalysis and WastageAnalysis.....	4
Figure 2. Correlation Matrix for Fraud Analysis data	8
Figure 3. ROC Chart for Random Forest and Decision Tree model	10
Figure 4. Precision-Recall Curve (Random Forest and Decision Tree models).....	10
Figure 5. Training Log Loss vs. Number of Rounds in XGboost model	11
Figure 6. ROC Curve for XGBoost model	12
Figure 7. Precision-Recall Curve for XGBoost model	12
Figure 8. Comparison of Validation Vs Test accuracy in XGBoost model.....	12

Table of Tables

Table 1. Statistical Summary and Correlation Matrix based on Fraud Analysis Data	8
Table 2. Potential Fraud cases summarization based on error code group by logic.....	8

Table 3. Summarization based on ClaimLabel_new column	8
Table 4. Random Forest and Decision Tree Classification Model Metrics	9
Table 5. XGBoost model metrics for Fraud analysis data	11

Chapter 1. Problem Statement and Scope of Dissertation

This dissertation aims to create a framework that applies data mining and machine learning techniques to combat fraud, waste, and abuse (FWA) in the processing of claims. Despite current safeguards, FWA persists in depleting resources, escalating costs, and compromising the trustworthiness of healthcare systems. The objective is to detect and thwart fraudulent actions, reduce superfluous expenditures stemming from waste, and curtail abuse within the claims processing infrastructure. Moreover, this framework can enable the company to enhance its revenue by reclaiming funds from providers through the re-evaluation and correction of improperly disbursed claims.

Scope of the project:

1. Framework Design

- a. Identify and collect relevant data sources, such as historical claims data, provider information, diagnosis codes, billing, and coverage information.
- b. Define the key variables and features that are indicative of FWA patterns using domain expertise and exploratory data analysis.
- c. Define a data pre-processing pipeline to clean, transform, and integrate the collected data.
- d. Explore and select appropriate data mining and machine learning techniques for FWA detection.

2. Development

- a. Implement machine learning algorithms, such as anomaly detection, classification, clustering, or predictive modelling, to identify potential cases of FWA. Ensure they are trained on a comprehensive dataset that includes known cases of FWA.
- b. Develop a fraud scoring system to prioritize suspicious claims based on their likelihood of being fraudulent.
- c. Implement rule-based systems to flag suspicious patterns and behaviors in claims processing.

3. Evaluation and Validation

- a. Evaluate the performance of the developed framework using appropriate evaluation metrics, such as precision, recall, F1-score, and accuracy.
- b. Validate the framework using a representative dataset, including both known fraudulent cases and normal claims.
- c. Fine-tune and optimize the models based on the evaluation results.

4. Likely Output

- a. A fraud detection framework capable of processing claims data and flagging potential cases of FWA.
- b. Identification of suspicious claims based on their likelihood of fraudulent activity.
- c. Insights and patterns discovered through data mining techniques, providing actionable information to prevent and mitigate FWA.

- d. Documentation and presentation of the project, including the methodology, results, and recommendations for implementation in real-world claims processing systems.

1. Development Methodology

i. Problem Understanding and Background:

- a. Gain a thorough understanding of the fraud, waste, and abuse challenges in claims processing.
- b. Review existing literature, research papers, and industry best practices related to fraud detection and prevention.
- c. Analyze the current claims processing system and identify potential gaps and vulnerabilities.
- d. Understand the available data sources, their quality, and limitations.
- e. Collaborate closely with business analysts and mentors and domain experts, to ensure alignment and gather valuable input throughout the development process.

ii. Data Acquisition and Pre-processing:

- a. Identify and collect relevant data sources, ensuring data privacy and compliance with regulations.
- b. Clean and preprocess the data by handling missing values, outliers, and inconsistencies.
- c. Conduct exploratory data analysis to gain insights into the data, identify patterns, and understand the characteristics of fraudulent claims.

iii. Feature Engineering and Selection:

- a. Engineer new features or transform existing features to capture relevant information for fraud, waste, and abuse detection.
- b. Use domain knowledge and expert input to identify informative features.
- c. Apply feature selection techniques, such as correlation analysis or feature importance ranking, to reduce dimensionality and improve model performance.

iv. Model Selection and Development:

- a. Evaluate and compare different machine learning algorithms suitable for fraud detection, such as anomaly detection, classification, or clustering.
- b. Train and fine-tune the selected models using a labelled dataset, including known fraudulent and normal claims.
- c. Optimize the models to improve their performance and generalization capabilities.

v. Rule-Based Systems:

- a. Develop rule-based systems to complement the machine learning models by capturing specific fraud patterns or behaviors that can be expressed as rules.
- b. Define and implement rules based on domain expertise and known fraud indicators.
- c. Integrate the rule-based systems with the machine learning models to enhance the overall fraud detection capabilities.

vi. Evaluation and Validation:

- a. Evaluate the developed framework using appropriate evaluation metrics, such as precision, recall, F1-score, and accuracy.
- b. Validate the framework's performance using a representative dataset, including both known fraudulent cases and normal claims.
- c. Conduct extensive testing and benchmarking against real-world scenarios and historical fraud cases.
- d. Refine and improve the framework based on the evaluation results and feedback and insights from the mentors.

vii. Documentation and Reporting:

- a. Document the entire development process, including methodology, algorithms used, and decisions made.
- b. Prepare a comprehensive project report, summarizing the methodology, results, and recommendations for implementation.

Justification of Suitability and Adaptation:

- a. Collaboration with subject matter experts and mentors ensures alignment with domain-specific knowledge and requirements, enhancing the framework's effectiveness in tackling fraud, waste, and abuse in claims processing.
- b. The inclusion of rule-based systems complements the machine learning models, incorporating specific fraud patterns and behaviors that can be captured as rules.
- c. The methodology accounts for the availability and quality of data, as well as the need for feature engineering and selection to enhance the models' performance and interpretability.

Chapter 2: Details on the Progress

2.1. Data Acquisition and Pre-processing:

- Collected 2 sets of claims data – one for Waste and Abuse analysis and another for Fraud claim analysis.
- Created a local database “rawData” in Microsoft SQL Server

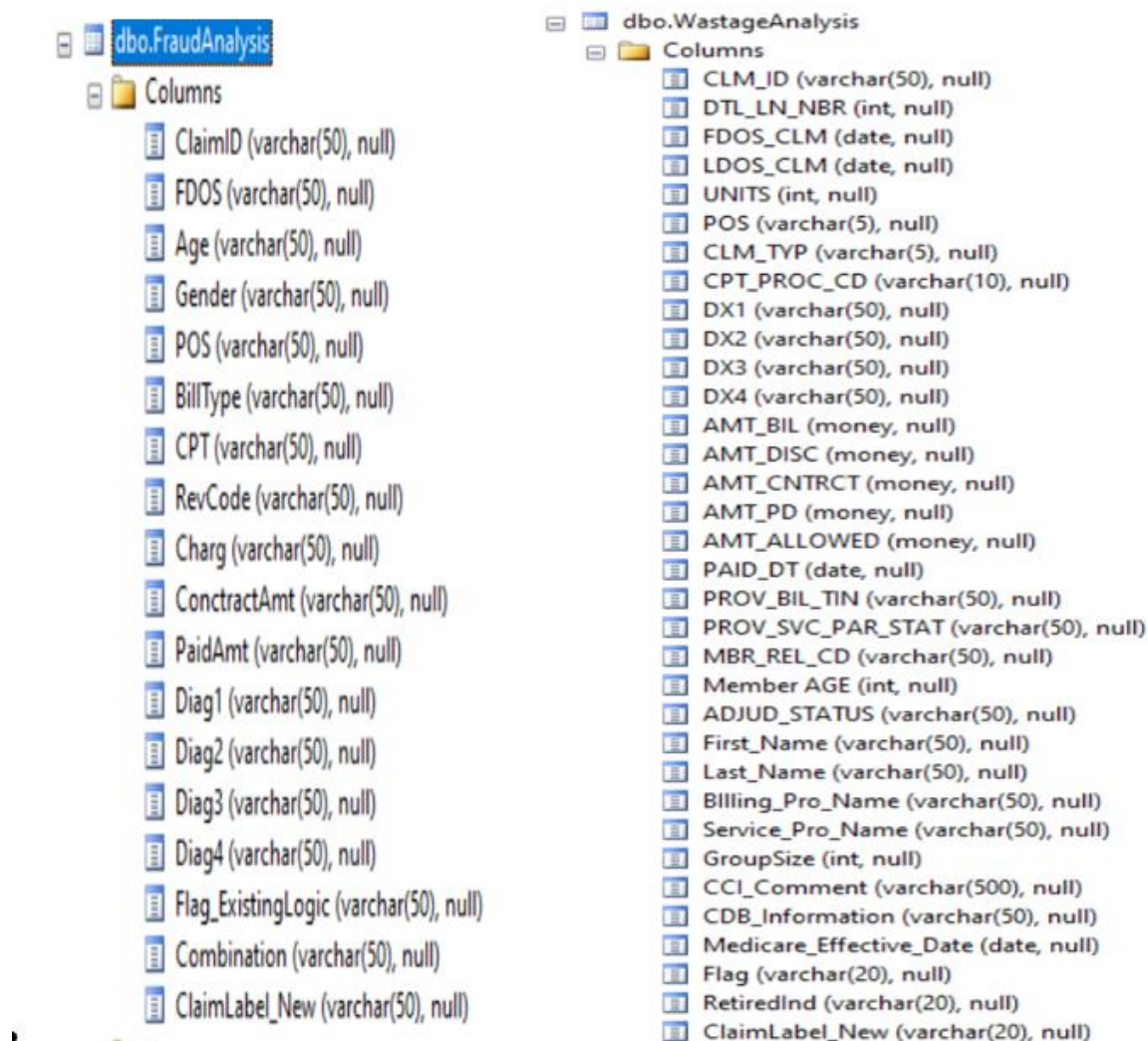


Figure 1. Tables and Columns for FraudAnalysis and WastageAnalysis

2.1.1 Pre-processing and Analysis on “Wastage Analysis” dataset

- Created a table called “WastageAnalysis” using SQL commands and defined the data type of each of the columns so that waste and abuse data can be imported to the table without any issue.
- Imported data into the table “WastageAnalysis.”
- Did some data de-identification / masking of PHI / PII information like First Name, Last Name, Billing Provider Name, Service Provider name, Claim ids. I used some commands

to strip or truncate some of the characters from these columns so that they can't be identified uniquely. Also, the claim ids were updated to make them some random but alpha numeric string.

- Based on the interactions with business analysts and subject matter experts working with Operations team, what I understood is there are multiple glitches within multiple applications in a claims adjudication data flow. Sometimes the right updates are also not cascaded to the downstream. Overall multiple updates or comments in the system are still done manually after a call to the provider or member. Those comments are updated in CCI_Comment column for the respective claims. Updates from CDB (Consumer data base containing member demographic information) are available in CDB_Information column. During the data analysis and research, it was concluded that many times this information are not in sync with the actual status of the member, or the coverage he/she has. Hence, discrepancy is observed in the processing of Medicare claims.
- If a member has Medicare coverage and if he/she satisfies other criteria for Medicare being primary coverage, then the claim paid by United Healthcare need to be reversed and the amount would be recovered from the provider. Based on more such information, I have gathered the below logic which can be used for Medicare claims processing.

Logic I used to set the Flag for claim classification : -

1. If Age \geq 65, Then member is retired. If Member does not have Medicare , then Flag is set as Correct which means UHG is the primary payer.
2. If Age \geq 65 and CCI_comment says = 'Medicare is Primary due to Retired' and CDB_Information = 'Yes Member has Medicare', and Medicare_effective_date < FDOS_CLM (First Date of Service of the Claim) then Flag is set as 'In-Correct'
3. If GroupSize < 20 and Age \geq 65 is 1 (True) and CCI_Comment says "Medicare is primary since group size < 20" , Then Flag is set as "In-Correct"
4. If Age < 65 and GroupSize < 20 and CCI_Comment says "Medicare is Primary due to Retired" and CDB_Information says "Yes-Member has Medicare", Then Flag is set as "Correct"
5. If Age < 65 and GroupSize > 20 and CDB_Information says "No-Member has no Medicare", Then Flag is set as "Correct".
6. If Age < 65 and GroupSize < 20 , Then set Flag as "Correct".

Sample data collected in WastageAnalysis database categorized into Correct or Incorrect, which means 12-13% of the sample data needs to be re-processed and the overpayment amount would have to be recovered.

Flag	ClaimRecordCount
Correct	87094
In-Correct	12875

2.1.2 Pre-processing and Analysis on “Fraud Analysis” dataset

- Similarly created a table called “FraudAnalysis” using SQL commands and defined the data type as required for the dataset.
- Did data de-identification / masking of PHI / PII columns like ClaimID
- Removed the un-used or other columns like First Name, Last Name, Provider details which are not required for my analysis.
- In the “FraudAnalysis” table, the column called “Flag_ExistingLogic” has 6 values like InPOS, InProc, InEM, InGen, InDiag, InOBs or null. Please refer to the list of abbreviations / acronyms to know the meaning of each of these Flags. Basically, this gives type of errors or flags the claims with an error code as per the existing logic to identify potential fraudulent claim.
- Now, using SQL commands and based on Diag1, Diag2, CPT, Rev Code, BillType, POS, Age and Gender, the “ClaimLabel_New” values have been set. This is as per the new logic gathered from the interactions with Operations team and domain experts.
- To find the top Diagnosis 1 and Diagnosis 2 codes for which maximum claims have come, I wrote the below SQL query.

```
select diag1, diag2, Flag_ExistingLogic, count(*) as n, sum(PaidAmt) as paidAmt from
FraudAnalysis where Flag_ExistingLogic <> 'NULL' group by diag1, diag2,
Flag_ExistingLogic order by count(*) desc
```

- To find the top Diagnosis 1 and diagnosis 2 codes for which maximum amounts have been paid as part of claim adjudication, I wrote the below SQL query.

```
select diag1, diag2, flag, count(*) as n, sum(PaidAmt) as paidAmt from FraudAnalysis
where Flag_ExistingLogic <> 'NULL' group by diag1, diag2, Flag_ExistingLogic order
by sum(PaidAmt) desc
```

- To find the claim volume count and paid amount for each of the Flags, I wrote the below query.

```
select flag, count(*) as VolumeCount, Sum(PaidAmt) as [Paid Amount] from
FraudAnalysis group by Flag_ExistingLogic
```

2.2 Model Development:

After initial analysis in SQL server database and defining the rule sets to classify a claim as Fraudulent or normal, I exported the data to csv file and connected to that dataset using python code.

To develop any prediction model after the data collection process, we normally follow the below process.

a) Data Pre-processing

- Cleaning: Remove or impute missing values.
- Feature Selection : Choose relevant features that could indicate fraudulent behavior.

Features relevant to our model development would be:

- Age
 - Gender
 - Diag1
 - Diag2
 - POS
 - CPT
- Feature Engineering: Create new features from existing data to improve model performance.
 - Normalization / Standardization: Scaled the features to treat all variables equally. Standard Scaler standardizes the dataset's features by removing the mean and scaling to unit variance, which is a requirement for the optimal performance of machine learning algorithms, including PCA.
 - Frequency Encoding: In this method, we replace each category with the frequency of its occurrence in the dataset. This can help the model understand the prevalence of each code without increasing dimensionality. By doing frequency encoding, we can capture the importance of category frequencies which might be predictive of the target variable.
In this Fraud data analysis, I did the frequency encoding on Diag1, Diag2 and CPT. As a result, I have new columns namely Diag1_Freq, Diag2_Freq, CPT_Freq.
 - Binary Encoding: Convert categories into binary columns, but with a logarithmic reduction in dimensionality compared to one-hot encoding. For e.g. Gender can be encoded as "Gender_M" with values as 0 or 1.

b) Exploratory Data Analysis (EDA)

- EDA is done to identify any patterns, anomalies, or relationships that could indicate fraud.

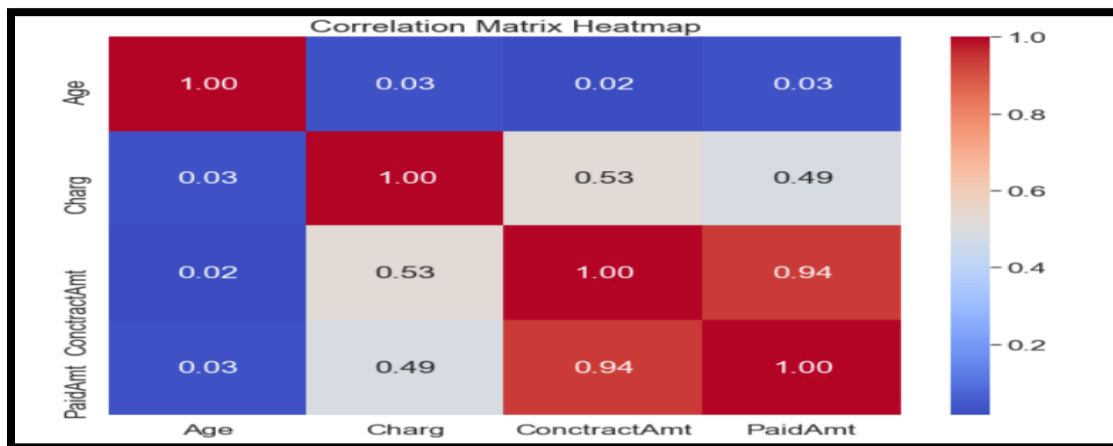


Figure 2. Correlation Matrix for Fraud Analysis data

Table 1. Statistical Summary and Correlation Matrix based on Fraud Analysis Data

Descriptive Statistics:				
	Age	Charg	ConctractAmt	PaidAmt
count	99870.000000	99870.000000	99870.000000	99870.000000
mean	39.604376	535.899510	250.984277	166.312725
std	16.152345	1162.175392	1036.561216	878.891206
min	0.000000	-21961.770000	-21961.770000	-21961.770000
25%	28.000000	89.000000	0.000000	0.000000
50%	41.000000	171.000000	0.000000	0.000000
75%	52.000000	337.000000	55.360000	3.460000
max	97.000000	61536.300000	59515.680000	59015.680000

Correlation Matrix:				
	Age	Charg	ConctractAmt	PaidAmt
Age	1.000000	0.029258	0.015516	0.027260
Charg	0.029258	1.000000	0.528083	0.488079
ConctractAmt	0.015516	0.528083	1.000000	0.940586
PaidAmt	0.027260	0.488079	0.940586	1.000000

Table 2. Potential Fraud cases summarization based on error code group by logic

	Flag_ExistingLogic	VolumeCount	Paid Amount
0	InDiag	2200	158049.02
1	InEM	5111	8578252.85
2	InGen	631	59978.75
3	InOBs	8	4307.48
4	InPOS	4707	382657.55
5	InProc	6	346.84

Table 3. Summarization based on ClaimLabel_new column

	ClaimLabel_New	VolumeCount	Paid Amount
0	Fraudulent	19058	9266305.94
1	Normal	80812	7343345.92

c) Splitting the Dataset

- Dividing the dataset into training and testing sets to evaluate model's performance.

d) Model Selection

- Selected three Machine learning algorithms (Decision Tree, Random Forest, XGBoost) for my dataset. The choice of Random Forest, Decision Trees and XGBoost stems from the fact that they have proven efficacy in handling complex, nonlinear relationships inherent in healthcare data.

- i. Random Forest does well in handling high-dimensional datasets with both numerical and categorical features, providing robustness against overfitting and capturing interactions among variables.
- ii. XGBoost or Extreme Gradient Boosting, includes regularization term in its objective function, which helps to avoid overfitting. It is designed to be highly scalable, efficient, and portable. It implements parallel processing and can handle the missing data. It allows the user to run a cross validation at each iteration of the boosting process. This type of model will split up to the max-depth level specified and then start pruning the tree backwards and remove the split beyond which there is no positive gain.
- iii. Decision Trees offer interpretability and simplicity, making them valuable for understanding the underlying decision-making process of the model.

e) Model Training

- Train the model using the training dataset and the selected algorithms.

f) Model Evaluation

- Testing of the model to evaluate the performance.
- Metrics such as accuracy, precision, recall, F1-score, and the area under the ROC curve (AUC) are important for evaluating a fraud detection model.

Table 4. Random Forest and Decision Tree Classification Model Metrics

Random Forest Classifier:					
	precision	recall	f1-score	support	
0	0.78	0.64	0.71	3795	
1	0.92	0.96	0.94	16179	
micro avg	0.90	0.90	0.90	19974	
macro avg	0.85	0.80	0.82	19974	
weighted avg	0.89	0.90	0.89	19974	
Accuracy: 0.8985180734955442					
Confusion Matrix:					
[[2446 1349]					
[678 15501]]					
Decision Tree Classifier:					
	precision	recall	f1-score	support	
0	0.62	0.66	0.64	3795	
1	0.92	0.91	0.91	16179	
micro avg	0.86	0.86	0.86	19974	
macro avg	0.77	0.78	0.78	19974	
weighted avg	0.86	0.86	0.86	19974	
Accuracy: 0.8599178932612396					
Confusion Matrix:					
[[2499 1296]					
[1502 14677]]					

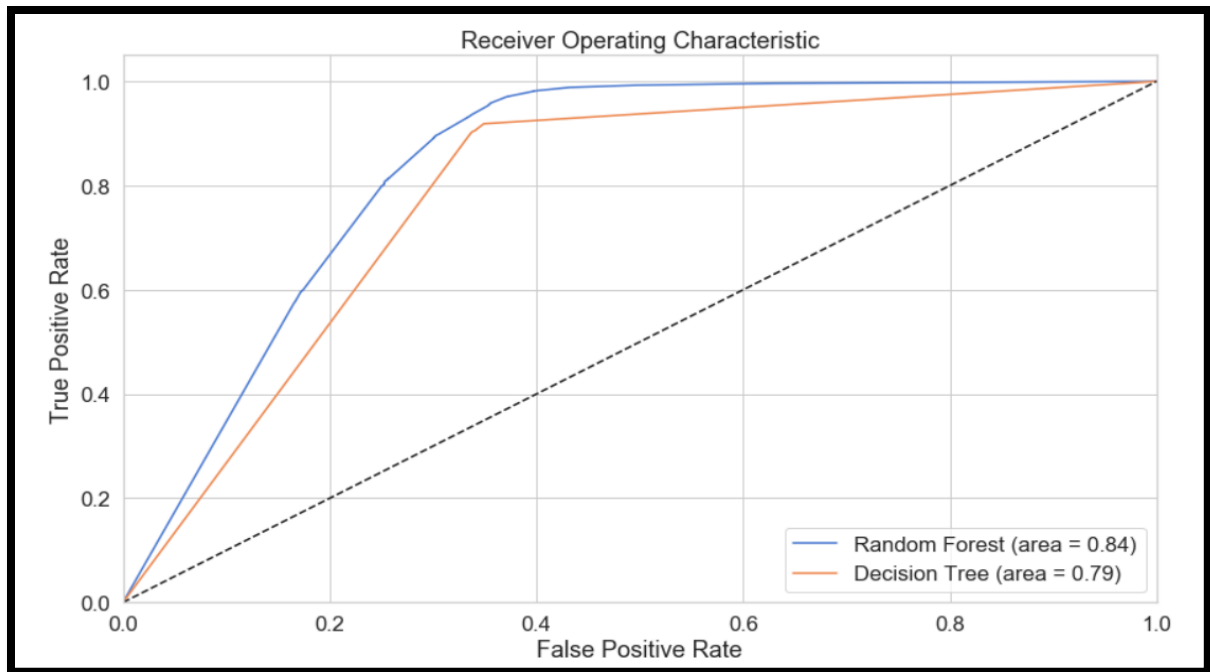


Figure 3. ROC Chart for Random Forest and Decision Tree model

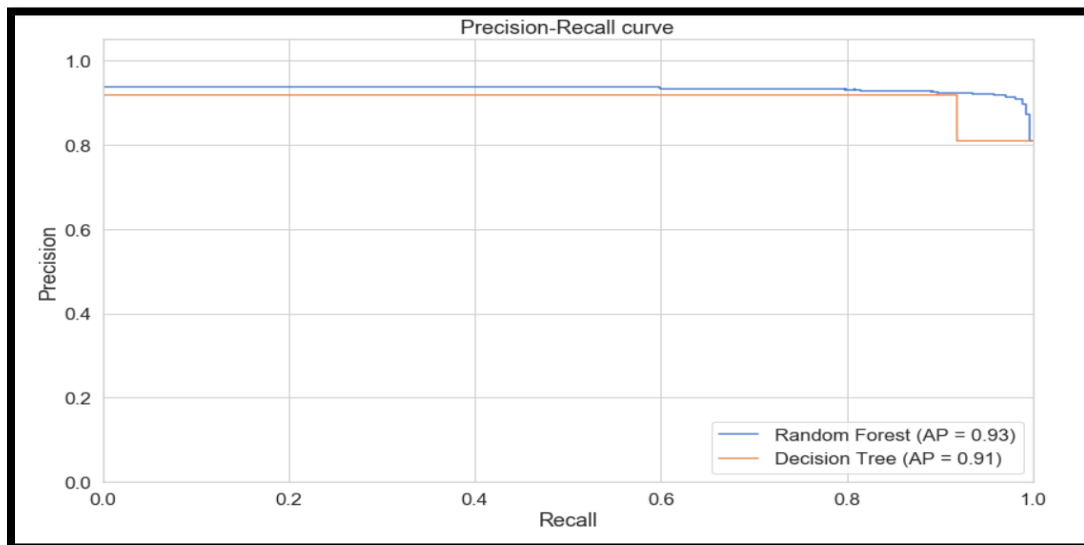


Figure 4. Precision-Recall Curve (Random Forest and Decision Tree models)



Figure 5. Training Log Loss vs. Number of Rounds in XGboost model

Table 5. XGBoost model metrics for Fraud analysis data

Class Index	Class Name				
0	Fraudulent				
1	Normal				
Test Accuracy: 0.92					
Classification Report for Test Data using XGBoost model:					
		precision	recall	f1-score	support
0		0.95	0.61	0.74	3795
1		0.91	0.99	0.95	16179
micro avg		0.92	0.92	0.92	19974
macro avg		0.93	0.80	0.85	19974
weighted avg		0.92	0.92	0.91	19974
Validation Accuracy: 0.92					
Classification Report for Validation Data using XGBoost model:					
		precision	recall	f1-score	support
0		0.94	0.61	0.74	2408
1		0.92	0.99	0.95	10376
micro avg		0.92	0.92	0.92	12784
macro avg		0.93	0.80	0.85	12784
weighted avg		0.92	0.92	0.91	12784

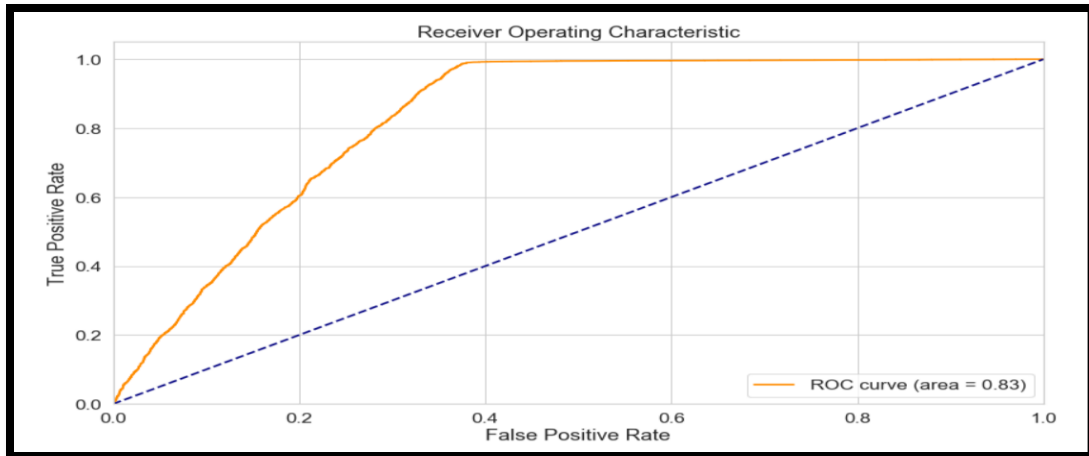


Figure 6. ROC Curve for XGBoost model

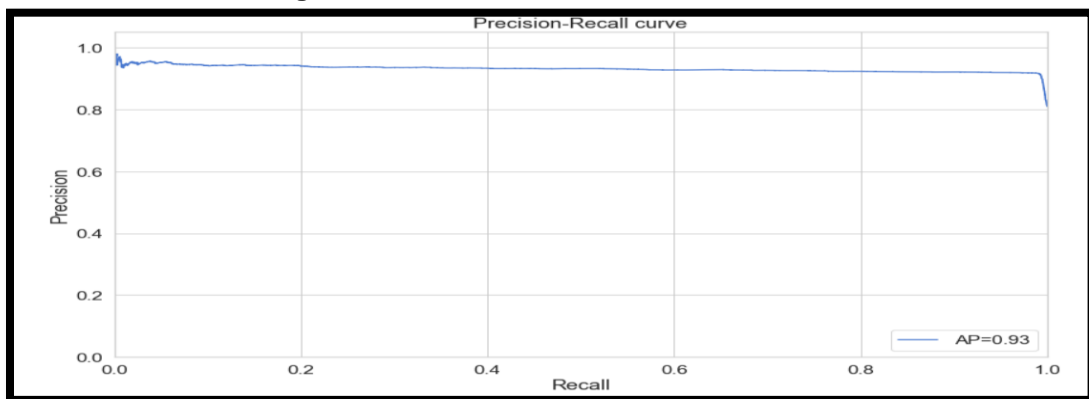


Figure 7. Precision-Recall Curve for XGBoost model

g) Explore on the Hyperparameter Tuning, if possible, to improve performance.

h) Model Validation

- Perform cross validation to ensure model's robustness.
- Further divided the test data set into Test and validation data set and displayed the results of the models.

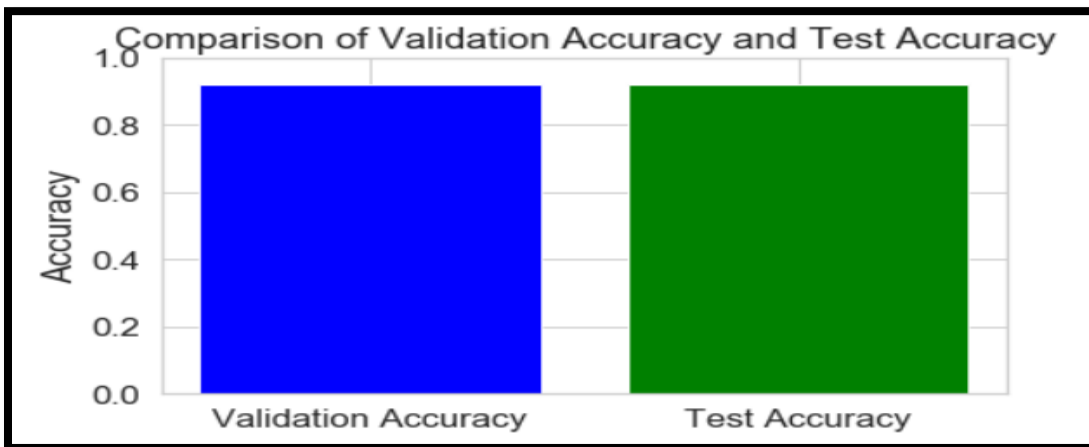


Figure 8. Comparison of Validation Vs Test accuracy in XGBoost model

i) Deployment

- Deploy the model into a production environment where it can start predicting fraud on new claims data. This is not in scope for this dissertation and may be considered in future if this POC is selected for deployment.

j) Monitoring and Updating

- Not in scope for this dissertation

Chapter-3: What is remaining and plan to complete it?

- More analysis on Waste Analysis and Fraud Analysis dataset
- Data Visualization using power BI.
- Model testing and refinement if possible
- Assess model performance using different metrics.
- Compare against existing process if any
- Project documentation and finalization
- Report preparation

List of abbreviations / Acronyms

- 1) FDOS – Date of Service
- 2) POS – Place of Service
- 3) RevCode – Revenue Code
- 4) PaidAmt – Paid Amount
- 5) Diag1 / Diag2 / Diag3 / Diag4 – Diagnosis codes
- 6) InPOS - Incorrect place of service
- 7) InProc - Incorrect procedure code
- 8) InEM - Incorrect emergency
- 9) InGen - Incorrect Gender
- 10) InDiag - Incorrect diagnosis code
- 11) InOBs - Incorrect Observation service
- 12) SQL – Structured Query Language
- 13) PHI – Personal Health Identifier
- 14) PII – Personal Identifiable Information
- 15) FWA – Fraud Waste and Abuse
- 16) TP – True Positive
- 17) FP – False Positive
- 18) TN – True Negative
- 19) FN – False Negative
- 20) CDB – Consumer Database
- 21) CCI – Claims Comment Information

Summary

In the precarious landscape of healthcare finance, a single substantial claim can precipitate a member's financial collapse. This dissertation underscores the imperative of discerning fraudulent claims without inadvertently denying legitimate ones due to flawed algorithms or review processes. While the primary aim is to devise machine learning models that flag potential fraud, thereby bolstering organizational revenue, paramount consideration is given to maintaining the integrity of customer and provider experiences. The organization's reputation hinges on transparent policy enforcement and claim adjudication, necessitating a robust redressal mechanism to safeguard genuine customers from undue care denial.

This dissertation presents a proof of concept for harnessing data science methodologies to detect fraudulent claims across diverse datasets. Current practices within Optum lack the integration of data mining or predictive analytics to preempt fraud, waste, or abuse. Predominantly, a rule-based system channels claims to operations for manual scrutiny, a method fraught with inefficiencies and susceptibility to error. This process not only burdens agents with excessive review of valid claims but also incurs unnecessary operational costs.

The crux of this research is to refine the claim review process, targeting a reduction in the volume of claims requiring manual assessment. By pinpointing potential fraud with greater precision, the model aims to curtail operational expenses and enhance revenue recovery from overpayments. The goal is to strike a balance between fraud prevention and the provision of uninterrupted, legitimate care, thereby upholding the organization's fiduciary and ethical standards.

By leveraging data driven approaches and predictive modeling techniques, this dissertation intends to pave the way for more effective and efficient process to tackle the highly prevalent FWA cases in US healthcare industry. Integration of predictive models and SQL based rules with the existing claims processing and recovering systems existing in the United HealthCare will ensure that the interests of genuine providers, customers, government, and claims payers are safeguarded, eventually resulting in improvement in health outcomes.

In summary, achieving a balance between detecting actual fraud (high recall) and minimizing false positives (high precision) is crucial in insurance fraud detection. Regular model evaluation and fine-tuning are essential to optimize these metrics based on business priorities and risk tolerance.

Literature Surveys

1. **Big Data Fraud Detection Using Multiple Medicare Data Sources:** Herland, Khoshgoftaar, and Bauder (2018) focused on detecting Medicare fraud using multiple CMS datasets. They employed data processing techniques on four datasets and mapped real-world provider fraud labels using the List of Excluded Individuals and Entities (LEIE) from the Office of the Inspector General. Their exploratory analysis involved building and assessing three learners on each dataset. The results showed that the combined dataset with the Logistic Regression learner yielded the best overall score, indicating the effectiveness of using combined datasets for detecting fraudulent behavior.
2. **Approaches for Identifying U.S. Medicare Fraud in Provider Claims Data:** In a subsequent study, Herland, Bauder, and Khoshgoftaar (2020) proposed an approach to predict a physician's expected specialty based on the type and number of procedures performed. They tested and assessed several new approaches to improve the detection of U.S. Medicare Part B provider fraud. Their results indicated that their proposed improvement strategies had mixed results over the selected Logistic Regression baseline model's fraud detection performance.
3. **Predictive Models to Detect Medicare Fraud, Waste, and Abuse:** The SMU Data Science Review (2023) discussed the development of a new model utilizing machine learning to generalize the patterns of fraud, waste, and abuse in Medicare. The model was trained on an Isolation Forest algorithm using previously fraudulent behavior. The results indicated anomalous instances occurring in 0.2% of all analyzed claims, demonstrating the predictive ability of machine learning models to detect Fraud, Waste, and Abuse (FWA).
4. **How Agencies Can Combat Fraud, Waste, and Abuse:** Cloudera discussed how technology leaders have developed a holistic solution to combat fraud, waste, and abuse by combining powerful processing capabilities with robust analytical tools
5. **The Health Care Fraud and Abuse Control Program:** The Centers for Medicare & Medicaid Services (CMS) highlighted the Health Care Fraud and Abuse Control (HCFAC) Program, which has been at the forefront of the fight against health care fraud, waste, and abuse since its inception in 1997.
6. **Fraud, Waste, and Abuse Toolkit—Health Care Fraud and Program Integrity: An Overview for Providers Booklet:** This booklet addresses common types of Medicaid fraud, waste, and abuse so that providers may recognize, report, and prevent them. It also discusses some of the program integrity measures against such activities.
7. **Best Practices for Employers Mitigating Health Care Fraud:** Moss Adams discussed the impact of health care fraud on employers and provided best practices for mitigating such fraud.

Appendices

Definition of the commonly used metrics to evaluate the model performance:

Some common **metrics** to analyze healthcare insurance data for Fraud detection are as follows.

1. True Positive (TP):

- In the context of insurance fraud detection, a true positive represents a transaction that the model correctly identifies as fraudulent. It means the model flagged a potentially fraudulent case, and upon investigation, it was indeed fraudulent.
- Significance: High TP rate ensures that actual fraud cases are detected, leading to timely action (e.g., blocking the transaction, investigating further, or alerting authorities).

2. False Positive (FP):

- A false positive occurs when the model incorrectly classifies a legitimate transaction as fraudulent. It raises a false alarm, potentially inconveniencing the customer.
- Significance: While minimizing FP is essential, too many false alarms can strain customer trust and operational efficiency. Balancing precision (low FP) and recall (high TP) is crucial.

3. True Negative (TN):

- In fraud detection, a true negative represents a legitimate transaction correctly identified as non-fraudulent. These are the “normal” transactions that don’t raise suspicion.
- Significance: High TN rate ensures that genuine transactions proceed smoothly without unnecessary scrutiny.

4. False Negative (FN):

- A false negative occurs when the model fails to detect a fraudulent transaction. It means the system missed a potentially fraudulent case.
- Significance: Minimizing FN is critical because missing actual fraud can lead to financial losses, damage to reputation, and legal consequences.

5. Precision:

- Precision (Positive Predictive Value) measures the proportion of flagged fraud cases that are genuinely fraudulent. Precision is calculated as $TP / (TP + FP)$

- Significance: High precision ensures that when the model raises an alert, it is likely a true fraud case. However, overly aggressive models may sacrifice recall (miss actual fraud) to achieve high precision.

6. Recall (Sensitivity):

- Recall measures the proportion of actual fraud cases correctly identified by the model. Recall is calculated as $TP / (TP + FN)$.
- Similarly, Fall-out (known as False Positive Rate) is calculated $FP / (FP + TN)$
- Significance: High recall ensures that the model doesn't miss actual fraud. However, overly sensitive models may generate more false positives.

7. Specificity (True Negative Rate):

- Specificity measures the proportion of actual non-fraudulent cases correctly identified by the model. Specificity is calculated as $TN / (TN + FP)$
- Significance: High specificity ensures that legitimate transactions are not unnecessarily flagged as fraud. It complements recall.

8. F1 Score:

- The F1 score balances precision and recall. It considers both false positives and false negatives. F1 Score is calculated as $2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$
- Significance: F1 score helps strike a balance between identifying actual fraud (recall) and minimizing false alarms (precision).

9. Accuracy:

- Accuracy measures overall correctness (both TP and TN) of the model's predictions. Accuracy is calculated as $(TP + TN) / (TP + TN + FP + FN)$
- Significance: While accuracy is essential, it can be misleading in imbalanced datasets (where non-fraudulent transactions significantly outnumber fraud cases). Focusing on precision, recall, and F1 score is often more informative.

Checklist of items for the Dissertation Report

This checklist is to be duly completed, verified and signed by the student.

1.	Is the final report neatly formatted with all the elements required for a technical Report?	Yes
2.	Is the Title page in proper format?	Yes
3.	(a) Is the Certificate from the Supervisor in proper format? (b) Has it been signed by the Supervisor?	Yes Yes
4.	Is the Abstract included in the report properly written within one page?	Yes
5.	Is the title of your report appropriate? The title should be adequately descriptive, precise and must reflect scope of the actual work done. Uncommon abbreviations / Acronyms should not be used in the title	Yes
6.	Have you included the List of abbreviations / Acronyms?	Yes
7.	Does the Report contain a summary of the literature survey?	Yes
8.	Does the Table of Contents include page numbers? (i). Are the Pages numbered properly? (Ch. 1 should start on Page # 1) (ii). Are the Figures numbered properly? (Figure Numbers and Figure Titles should be at the bottom of the figures) (iii). Are the Tables numbered properly? (Table Numbers and Table Titles should be at the top of the tables) (iv). Are the Captions for the Figures and Tables proper? (v). Are the Appendices numbered properly? Are their titles appropriate	Yes Yes Yes Yes Yes Yes
9.	Is the conclusion of the Report based on discussion of the work?	Yes
10.	Are References or Bibliography given at the end of the Report? Have the References been cited properly inside the text of the Report? Are all the references cited in the body of the report	N/A N/A N/A
11.	Is the report format and content according to the guidelines? The report should not be a mere printout of a Power Point Presentation, or a user manual. Source code of software need not be included in the report.	Yes

Declaration by Student:

I certify that I have properly verified all the items in this checklist and ensure that the report is in proper format as specified in the course handout.



Signature of the Student
Name: Soumya Ranjan Pati
BITS ID:2022OG04006