Prescription Fraud Analysis

Obay Baradie

500888472

2024-10-27

Ceni Babaoglu

Ryerson University

Contents

Abstract

In the US, Medicare plays a critical role in the healthcare system by administering and overseeing drug prescription coverage for beneficiaries nationwide. However, with the evident rise in healthcare costs and abnormalities in drug consumption, reports and suspicions of fraud within this industry have been noted. This makes it necessary to dive deeper into precision administration and fund allocation as it affects millions of patients nationwide. This research aims to analyze for anomalies within the practices that can point to fraudulent activity within the industry. Anomalies in drug spending may include such behaviour as phantom billing, duplicate claims, or over-prescription of some medications. These behaviours can harm patients, companies, and several other parties involved. The effects can range from financial disciplines to much larger physical ones. Also, these fraudulent practices can increase healthcare costs and lower the effectiveness of the drug administration system. Detecting these anomalies is crucial for maintaining wellness and integrity within the public healthcare system.

The research question driving this analysis project is: How can anomalies in Medicare drug spending help conquer fraudulent overprescriptions in the healthcare industry?

An analysis of prescription spending data will be used to:

1. Identify abnormal spending patterns among Medicare parties and drugs
2. Recognize those drugs with abnormally high/low spending, claims, or usage

3. Provide insights into how these anomalies can be used to detect fraud, such as phantom billing or over-prescription.

By analyzing spending patterns and anomalies, we aim to discover outlier trends to help policymakers and healthcare companies make informed decisions.

Data:

The resources for this analysis will be a comprehensive list of data sets that meticulously track the administration of prescription drugs and several key features, such as Brand Name, Dosage units, Total Claims, and Average spending per drug. This comprehensive approach ensures that every stone is turned on in our quest to uncover and combat fraudulent overprescription. Another dataset that will be explored throughout this project is the Medicare hospital spending by claim. This dataset is rich in information regarding the expenditure for claims in many hospitals in the US. Specific features like claim type, average costs, and percent of spending in hospitals can aid in analyzing and discovering trends within hospital claim spend.

Techniques and Tools:

The project will employ clustering algorithms to detect outliers in the drug/ claim spending data and conduct an anomaly analysis. A random forest model will be chosen to develop decision trees to identify patterns and distinguish between outlier (fraudulent) claims and normal ones. Each decision tree in the forest learns different aspects of the data, focusing on specific features (e.g., claim amount, provider type, drug name, etc.) that could indicate outlier status (fraud).

In terms of tools used, this project will use Python with analysis libraries like Pandas, Numpy, and Matplotlib. Additionally, Tableau will generate more prominent visualizations and dashboards for presentation.

Conclusion:

This project aims to detect abnormal Medicare drug spending data patterns that may indicate fraudulent activity by applying clustering techniques. The insights from this analysis can help healthcare providers and policymakers address inefficiencies, reduce costs, and maintain the integrity of the Medicare Part D program.

Introduction

Health insurance fraud is a significant problem that continues to cause tremendous monetary losses in the healthcare industry. This phenomenon can affect every individual and party involved in the healthcare transaction. Moreover, this can range from a beneficiary to a well-renowned institution. The National Health Care anti-fraud Association (NHCAA) estimates that the fraud-related financial loss is roughly tens of billions of dollars in the US alone (Haque & Tozal). Recent figures estimate that loss due to healthcare fraud is about 68 billion annually. More specifically, medicare spending increased from $471 billion to $798 billion in a decade from 2009 to 2019 (Settipalli & Ganagdharan, 2022). While these figures can be partially explained by general economic factors such as inflation, another prominent explanatory reason is the evident surge in healthcare claims costs. In countries like the USA and India, it is estimated that the number of fraudulent claims in healthcare is approximately 15% of total claims (Settipalli & Ganagdharan, 2022). Fraudulent activities can include a diverse list of types. However, the most common types typically involve Making false diagnoses to justify unnecessary procedures, billing for higher priced procedures (upcoding), billing for each step of the method as if it were a separate procedure (unbundling), and misrepresentation of non-covered treatments as medically necessary, especially for cosmetic procedures (Haque & Tozal). When looking at fraud in the prescription space, it is evident that the most common approach to claim fraud is through overprescription and unnecessary prescription of drugs for patients. According to the US Ministry of Justice, prescription fraud in certain U.S. states is rivalling traditional street drug pedalling. These fraudulent prescription instances are tragic because of the magnitude of harm they can cause patients often seeking genuine treatment. According to the New Hampshire Department of Health and Human Services, more than 400 people died as a result of a drug overdose in 2015, almost 2.5 times more than the overdoses in 2011 (Zafari & Ekin, 2018). Instances like this can be explained by the fraudulence of provider overprescription, especially when the client’s needs are exaggerated. This ties into over-billing, where some physicians will target specific clients by giving them false diagnoses to increase their overall medical bill. This will eventually be associated with a much larger claim that they can benefit from their insurance provider. Other fraudulent behaviours include altering prescriptions, claiming reimbursement for non-provided treatments, and generating ghost patients (Johnson & Nagarur, 2015). On the patient, however, fraud behaviours commonly occur as well. They, for instance, may provide incorrect medical history, false demographic information, or downplay their financial status to receive better coverage. Some specific examples include submitting claims for ineligible dependents, filing for unreceived prescriptions, or identity theft (Johnson & Nagarur, 2015). Similarly to health provider fraud, patients can commit their share of fraud, resulting in detrimental financial losses for physicians and providers involved in the claim. Any prescription fraud is associated with heavy negative connotations. Decades of research and studies have been aimed at finding statistical ways to analyze and detect insurance fraud in the medical industry—past research involved methods and employment for models designed for outlier detection. Moreover, model parameters were defined, and an objective was set based on the research questions and data. Much of the past work on the subject can be categorized into supervised or unsupervised machine learning. Supervised machine learning refers to the algorithm that learns from a labelled dataset, and each input comes with a corresponding output or target variable. Common examples include regression and classification algorithms. On the other hand, an unsupervised approach is when an algorithm learns from unlabeled data without pre-existing output labels. Clustering and dimensionality reductions are common examples. There is abundant research on fraud detection in healthcare, and there is a need for both types of learning algorithms. For this research project, a supervised learning approach will be undertaken to factor in the explanatory variable present in the medicare part D dataset. Furthermore, a classification (Random forest) algorithm will be used to explain the class and features of a prescription claim deemed fraudulent. More specifically, the study will look at thousands of medications to determine which types/brands of medications are most affected by fraud and their implications on the wellness of the health industry.

Related Work

More than several studies in the past have taken on the responsibility of building a model to detect outliers or frauds in their respective industries. As mentioned, the distribution of supervised vs. unsupervised learning techniques is quite even, and both have produced significant results in the field. (Bayerstadler et. al) have used a multinomial model based on the Bayesian latent variable, predicting fraud and abuse probabilities. Their study used a systematic analytical approach to identify the fraudulent behaviour based on a “reporting factory” which considers the providers’ network and invoice properties. Additionally, a predictive scoring model was used to classify new invoices into three distinct categories: *unperformed services (fraud), unjustified services (abuse), or other billing issues (fraud/abuse).* The dataset contained over 100,000 manually reviewed cases and was extracted from an archive as part of a known insurance company in the Middle East. The sampling approach used B = 50 subsamples, with 75% of all fraud/abuse cases occurring in at least one of those subsamples. Furthermore, the model predicts the probability that an invoice belongs to one of the mentioned categories and assigns probabilities based on latent variables. The model demonstrated high predictive performance, with the latent variable approach proving useful in detecting fraud across the different categories. Another prominent model in the field was introduced by (Settipalli and Ganagdharan, 2022) when they developed their unsupervised multivariate analysis model, which was used for fraud detection in health insurance claims. Using medical claims obtained from CMS Part B in 2018, the model analyzed multivariate categorical data and continuous data in two stages: the first stage constructed a weighted multitree (WMT) for categorical data to observe provider behaviour trends. The second stage was designed to detect false claims using a univariate fraud detection model using density-based clustering. This approach proved advantageous for the researchers as it removed the need for labelled data. Additionally, with the WMTDBC model, the graph-based WMT avoids class overlaps, which ensures the separation of distinct claim groups. Overall, the model employed proved effective at anomaly detection, with the density-based clustering accurately identifying deviations in service counts and claimed amounts. The model significantly improved in reducing false alarms and increasing detection accuracy compared to traditional methods. (Johnson & Nagarur, 2015) proposed a claim fraud detection approach that used provider profiling. The researchers identified providers with high patient-encounter ratios, long patient visits, and excessive medication counts. for the distance calculation, the Mahalanobis distance was used because it yielded better segmentation than the Euclidean distance when there is a correlation among other variables. Additionally, the study used a density calculation to detect the providers practicing differently, which might have been missed by the distance calculation. The data comprised 878,691 claims from 2076 providers across the four specialties: otolaryngology, general practice, neurology, and ophthalmology. The study analyzed the claims data through six stages to determine the relevancy and accuracy. For example, after provider profiling, the data underwent a demographic screening, which later translated into a claim amount screening to identify overstated amounts. The proposed methodology was compared with unsupervised neural networks. Overall, the new methodology achieved a rating accuracy of 84-87% across different specialties, which means it also outperformed the neural networks. Computer-aided audit of pharmaceutical prescriptions was carried out by (Iyengar et al., 2013), where their analysis focused on four drug classes associated with fraud and abuse. The four drugs studied were Narcotics, Ataractics-tranquilizers, CNS stimulants, and Amphetamine Preparations. The methodology was driven by the rule list models that segment the input space into homogenous groups based on the behaviour of entities. Using their models, they calculated prescription rates for each segment. In addition, Likelihood Ratio Tests (LRT) were used to check whether a group's behaviour deviated from the norm. Finally, the scores were aggregated to rank the most suspicious entities. Monte Carlo simulations were used to test the statistical significance of these scores. The dataset contained roughly 600,000 claims associated with different drugs and brand names. The baseline models successfully predicted the expected prescription behaviours across different drug classes. Overall, the models identified entities with high and low rates of prescription fraud, and these findings aligned with known drug interactions and disease treatment patterns. Regarding an unsupervised approach, (Joudaki et al., 2015) aimed to improve physician claim fraud detection through a comprehensive data mining study. The data was collected from the provincial branch of the SSO in Iran, which includes 454 general physicians, specialists, and dentists. Later on, the entries were organized into two datasets. One included physician-level variables and indicators on 164 general physicians. The second data set included 474897 drug prescription claims. A hierarchical clustering method segmented physicians into groups suspected of fraud and abuse. the unsupervised approach incorporated Euclidean distance to find the optimal number of clusters using the silhouette coefficient. As a result, 2 clusters were developed and labelled as either healthy or suspect (fraud-related). Through the data mining, Thirteen indicators were developed. 2 related to cost issues, 4 to frequency and pattern of visits, and seven were related to prescription patterns. The higher the indicator's value, the greater the possibility of fraudulent behaviour. On the fraud detection front, physicians were categorized into 2 clusters. The suspect group (cluster 2) were more likely to have patients visiting more than once in a short period. Suspect physicians were 50 times more likely to have their prescriptions dispensed at high-cost pharmacies, and the cost of their prescriptions was about 60% more than others. (Shin et al., 2012) utilized a scoring model to detect abusive patterns in health claims. In South Korea, approximately 3.6% of the population is covered through medical aid programs funded by the general tax, while the remaining pertains to the National Health Insurance Corp (NHIC). Although the government continues to find ways to budget and decrease health spending, the health insurance figures continued to expand at an average annual rate of 16%. The model depicted two phases: one for scoring, where an apriori search was used to locate the top and low-ranked providers. The other was to segment the composite degree of anomaly (CDA) scores so reviewers could tailor responses toward abusive providers. After sorting providers, they were arranged into several groups for grade for the degree of anomaly (GDA) classification. The providers most likely to be abusive—the ones with a very high CDA score, GDA(4)—are segregated into two segments based on the difference in their abusive utilization patterns: the CI of total charges and the rate of prescribing antibiotics. The data from the HIRA included 45,000 records of 187 indicators for 28,066 clinics in general practice and 16 specialists. The study was split into two subsets, the third and fourth quarters of 2007 (Shin et al., 2012). Based on their results, The least significant indicators, with values less than 1, include the prescription rate of costly medications, the number of visits per claim, and the CIs of charges for CT, MRI, registration, and CMI. The claims in the 236 GDA(4) clinics (6%) revealed the most abusive utilization patterns and the payer could expect the largest gain by correcting their utilization behaviour. Lastly, a confusion matrix was developed, and its results show that the CDA model identified some providers (cells c and g) that the manual process missed, indicating that the model could be more effective in catching problematic providers. Overall, The model raises concerns about the effectiveness of the current manual process, especially in cases where providers not flagged for intervention by HIRA had high CDA scores and should have been reviewed. While (Joudaki et al., 2015) aimed to detect fraud instances through a data mining study, (Kumaraswamy et al., 2022) used the feature engineering approach to derive insights into their claims data. Their methodology involved data pre-processing, feature engineering, and feature extraction on the List of Individuals and Entities (LEIE) data set owned by Texas Health. The dataset had 30 variables that described each 2016 transaction from all Medicaid clients and pharmacies. Since their analysis focused on designing features for fraud prediction, the authors performed four separate experiments for validity. The first used aggregated features engineered from the prescription claims. Similar to the first, the second aggregated features but added a synthetic minority oversampling technique (SMOTE) to address the class imbalance. The third and fourth experiments incorporated the first ‘n’ principal components of PCA and a combination with SMOTE. In addition to the four experiments, logistic regression and a random forests model were tested on the engineered feature-based data matrix. In terms of findings, the PCA showed that a linear combination of 15 principal components explained 85% of the variance seen in the claims. The logistic regression and random forest models performed best (Kumaraswamy et al., 2022). They had an F1 score equal to 0.18 for the class of interest on the testing data and a weighted F1-score of 0.85 (logistic regression) and 0.88 (random forest). The final study analyzed (Zafari & Ekin, 2018) was motivated by prescription fraud abuse in the state of New Hampshire with an emphasis on identifying providers with unusual prescription rates, especially opioids. According to the New Hampshire Department of Health and Human Services (2017), more than 400 people in New Hampshire died as a result of a drug overdose in 2015, two and half times more than the overdoses in 2011. Additionally, this state has the second-highest per capita drug overdose deaths (34.3 per 100000) in the US. (Zafari & Ekin, 2018)). They wanted to construct a study that analyzed topic models with covariates to help determine the overprescribed drugs. This work uses the topic models to form benchmark groups based on prescription patterns. The outliers are then detected based on deviations from the benchmark distributions. We use the concentration function and distance-based measures to capture deviations from the expected behaviours. The study used the medicare part D prescriber data set for 2015. Five thousand seven hundred one medical providers from 88 specialties submitted over 5.43 million medicare part D claims in New Hampshire in 2015. Through the grouping of prescribed-drug pairs, it can be seen that specialties such as ‘interventional pain management’ are top prescribers of the opioid drug group. The researchers deduced that The structural topic models allow for flexibility by including provider-level covariates to capture prescription patterns better, although this comes with a computational cost. Also, The framework leverages concentration functions and various outlier measures to detect abnormal prescription behaviours, acting as a prescreen filter for identifying potential fraud cases.

Methodology:

Fraud claims analysis has been conducted several times, with a 50-50 distribution between supervised and unsupervised approaches. A quantitative research study will be conducted to analyze and determine claims fraud related to thousands of prescription drugs in the US. More specifically, the aim will be to uncover which drugs and their respective brands are most associated with insurance abuse. This finding will prove to be vital for the future and longevity of the US healthcare system. The study will utilize a dataset like (Zafari and Ekin's), Medicare Part D Spending by Drug. However, the observations will be taken from 2021-2022 to account for the recency bias. This dataset includes all the information on drugs patients generally administer, which are paid through the Medicare Part D Prescription Drug program. This data also displays spending information for drug manufacturers, denoted by the brand name feature. The Medicare dataset focuses on the mean spending per dosage and uses percentages to show the change in spending over time. This study will focus on the spending patterns in 2021 and 2022. Since the claims involve similar drugs of different dosages and strengths, the average spending per dosage unit is weighted to account for the variation in such factors. The overall brand name/generic name claim weighted spending per unit is calculated by first summarizing each drug to specific strength, form, route of administration, and manufacturer levels. For each unique level, spending is divided by the number of units and multiplied by its proportion of total claims so that claims volume becomes the weight (Medicare, 2024).

To develop a model that can effectively predict fraud within prescription claims data like Medicare Part D, the technique must be efficient and capable of handling high-dimensional and complex data. For this reason, a random forest model is utilized to carry out the analysis. Random forests are also excellent at dealing with non-linear relationships, which can be confirmed with the claim's dataset. Random forests are non-parametric, making them ideal for capturing the complex, non-linear relationships common in healthcare data due to patient behaviours, drug efficacy variations, and regional pricing differences. The statistical testing and modelling will be performed in Python using the Sklearn library. Here is a brief workflow on how a random forest might be used for this analysis:

Data Preparation and Feature Engineering:

- Clean the dataset and convert necessary columns, like monetary values, into numerical format.
- Identify columns with large amounts of missing data.

Balancing the Dataset:

- Check the balance between flagged and non-flagged claims and apply SMOTE or other resampling techniques if necessary.

Model Training:

- Split the dataset into training and testing sets.
- Train a random forest classifier with the "outlier flag" as the target variable and all other variables as predictors.

Feature Importance Analysis:

- Use feature importance to identify critical drivers of outlier behaviour, informing strategic adjustments in claims processing or prescription guidelines.

Evaluation:

- Assess the model using metrics like precision, recall, and F1-score, focusing on recall for outliers to ensure it correctly identifies high-cost or high-risk claims

Descriptive Statistics:

The Medicare Part D Spending By Drug 2022 is a dataset comprised of 13,889 records of drugs with their associated brand names. The features are split into 46 different variables, with the key variables being:

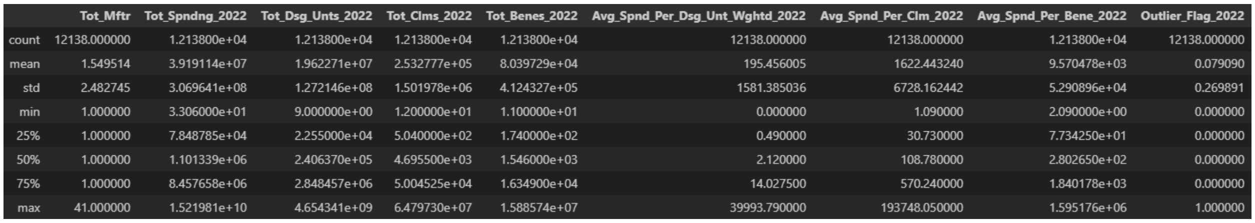
- Brnd_Name: Brand name of the drug.
- Gnrc_Name: Generic name of the drug.
- Tot_Mftr: Total number of manufacturers associated with the drug.
- Mftr_Name: Specific manufacturer name.
- Tot_Spndng_2018-2022: Total spending on the drug across years 2018-2022.
- Tot_Dsg_Unts_2018-2022: Total claims for each year.
- Tot_Benes_2018-2022: Total beneficiaries for each year.
- Avg_Spnd_Per_Dsg_Unt_Wghtd_2018-2022: Weighted average spending per dosage unit for each year.
- Outlier_Flag_2018-2022: Binary flag indicating potential fraud for each year.
- Chg_Avg_Spnd_Per_Dsg_Unt_21_22:Change in average spending per dosage unit between 2021 and 2022
- CAGR_Avg_Spnd_Per_Dsg_Unt_18_22: Compound annual average spending per dosage unit growth from 2018 to 2022.

It is worth noting that recency was heavily valued for this study to produce the most consistent results. In addition, missing values were abundant for the entries within the 2018-2021 columns, accounting for over 20% of values. For this reason, the columns relating to claims in 2018-2021 were negated, and the focus of this analysis will be on the information from 2021 and 2022.

Summary Statistics for numerical and categorical variables:

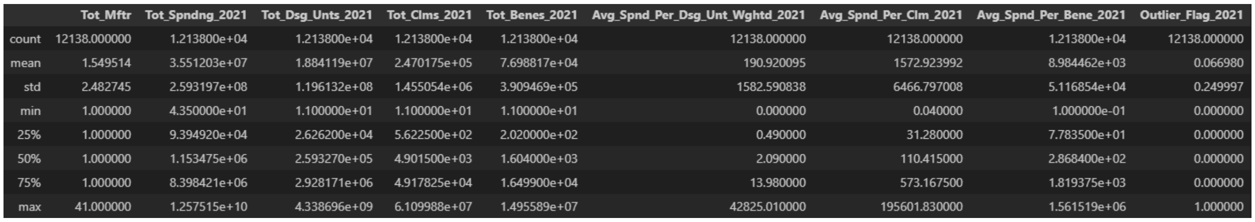
Below, the graphs for statistics are summarized. Tables are divided by year to display the difference in values and amounts.

Table 1.1 Summary Statistics 2021



	Tot_Mftr	Tot_Spndng_2022	Tot_Dsg_Unts_2022	Tot_Clms_2022	Tot_Benes_2022	Avg_Spnd_Per_Dsg_Unt_Wghtd_2022	Avg_Spnd_Per_Clm_2022	Avg_Spnd_Per_Bene_2022	Outlier_Flag_2022
count	12138.000000	1.213800e+04	1.213800e+04	1.213800e+04	1.213800e+04	12138.000000	12138.000000	1.213800e+04	12138.000000
mean	1.549514	3.919114e+07	1.962271e+07	2.532777e+05	8.039729e+04	195.456005	1622.443240	9.570478e+03	0.079090
std	2.482745	3.069641e+08	1.272146e+08	1.501978e+06	4.124327e+05	1581.385036	6728.162442	5.290896e+04	0.269891
min	1.000000	3.306000e+01	9.000000e+00	1.200000e+01	1.100000e+01	0.000000	1.090000	2.090000e+00	0.000000
25%	1.000000	7.848785e+04	2.255000e+04	5.040000e+02	1.740000e+02	0.490000	30.730000	7.734250e+01	0.000000
50%	1.000000	1.101339e+06	2.406370e+05	4.695500e+03	1.546000e+03	2.120000	108.780000	2.802650e+02	0.000000
75%	1.000000	8.457658e+06	2.848457e+06	5.004525e+04	1.634900e+04	14.027500	570.240000	1.840178e+03	0.000000
max	41.000000	1.521981e+10	4.654341e+09	6.479730e+07	1.588574e+07	39993.790000	193748.050000	1.595176e+06	1.000000

Table 1.2 Summary Statistics 2022



	Tot_Mftr	Tot_Spndng_2021	Tot_Dsg_Unts_2021	Tot_Clms_2021	Tot_Benes_2021	Avg_Spnd_Per_Dsg_Unt_Wghtd_2021	Avg_Spnd_Per_Clm_2021	Avg_Spnd_Per_Bene_2021	Outlier_Flag_2021
count	12138.000000	1.213800e+04	1.213800e+04	1.213800e+04	1.213800e+04	12138.000000	12138.000000	1.213800e+04	12138.000000
mean	1.549514	3.551203e+07	1.884119e+07	2.470175e+05	7.698817e+04	190.920095	1572.923992	8.984462e+03	0.066980
std	2.482745	2.593197e+08	1.196132e+08	1.455054e+06	3.909469e+05	1582.590838	6466.797008	5.116854e+04	0.249997
min	1.000000	4.350000e+01	1.100000e+01	1.100000e+01	1.100000e+01	0.000000	0.040000	1.000000e-01	0.000000
25%	1.000000	9.394920e+04	2.626200e+04	5.622500e+02	2.020000e+02	0.490000	31.280000	7.783500e+01	0.000000
50%	1.000000	1.153475e+06	2.593270e+05	4.901500e+03	1.604000e+03	2.090000	110.415000	2.868400e+02	0.000000
75%	1.000000	8.398421e+06	2.928171e+06	4.917825e+04	1.649900e+04	13.980000	573.167500	1.819375e+03	0.000000
max	41.000000	1.257515e+10	4.338696e+09	6.109988e+07	1.495589e+07	42825.010000	195601.830000	1.561519e+06	1.000000

Notable statistics include the means for both years, with an evident increase in the average spending per claim from 1572.92 to 1622.44 in 2022. This increase can partially be explained by the increased prices of drugs as well as the potential for fraud indicators. In addition, looking at the outlier flag variable, where 0 and 1 stand for no and yes, respectively, the overall mean of the dummy explanatory variable increased from 0.066 to 0.079, proving the overall increase in suspected fraud claims over the years. Although these figures offer insight into the fraud outlook, further analysis will be conducted to solidify these findings.

Outlier Analysis:

The Medicare Part D dataset was examined to identify potential fraud patterns, represented by the "Outlier Flag" variable across multiple years (2021-2022). the average outlier rates for each drug and its associated provider were analyzed, offering insights into drugs and manufacturers with higher instances of flagged claims.

On average, outlier rates were low across most drugs and providers, with only a tiny subset showing elevated flags. Drugs like *Abacavir*, represented by providers like *Aurobindo Pharm*, consistently exhibited low outlier rates across all years. This indicates a general trend of low fraud-related activity in the dataset; however, isolated cases may warrant deeper investigation, especially for drugs or providers showing atypically high outlier rates in specific years. Looking at the outlier_flag counts for both years below; it is noticeable that the outlier flags are significantly distributed on the opposing side. This means that over 80 observations in the dataset are associated with non-outlier (non-fraudulent) indicators. Because of this imbalance, the Synthetic Minority Over-Sampling Technique (SMOTE) will balance the class distribution. This will help the proposed random-fort model learn patterns in both classes more effectively, improving its sensitivity to fraud cases. Additionally, the random-forest model will provide more balanced data by applying SMOTE, which can reduce the likelihood of overfitting to the majority class. Overall, the study will greatly benefit from incorporating this technique, and further discussion will be included in the results section.

Table 1.4 Outlier Count 2021

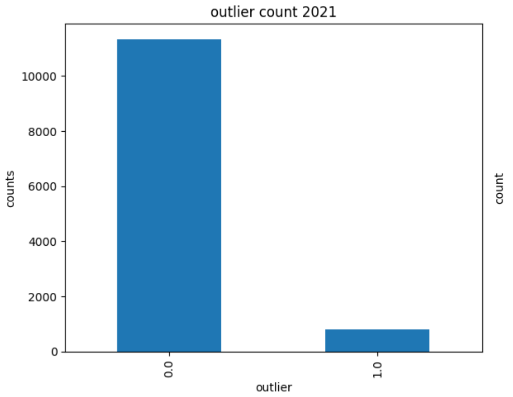
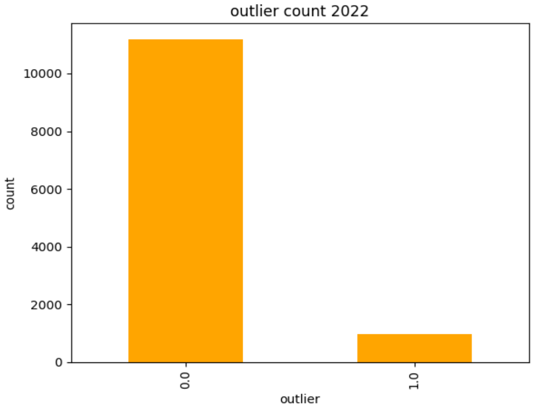


Table 1.3 Outlier Count 2022



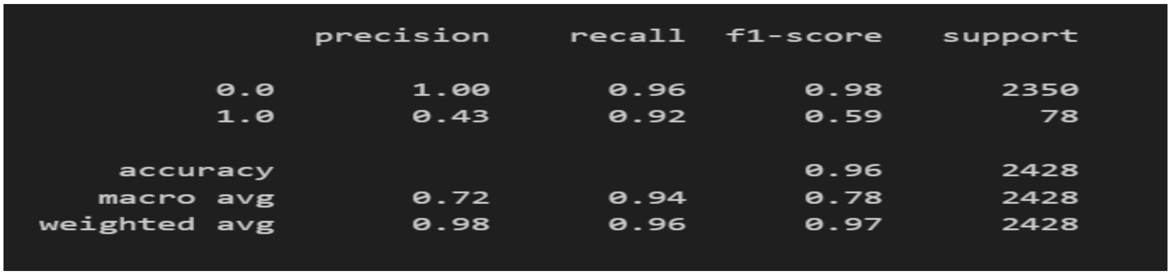
Results:

This analysis aimed to create a model that can accurately detect outliers within the prescription claim dataset using a random forest model. Although these outliers are not necessarily fraudulent, they can help researchers understand the characteristics of fraudulent prescription claims. In addition, by setting up a baseline model, continuous adjustments can improve the metrics and create a better model capable of detecting fraud. The study ran two different models for the years 2021 and 2022. Each year, an initial random forest was executed to evaluate the model's performance. After the initial results, SMOTE was carried out to attempt to improve the model by balancing the explanatory class. After SMOTE, an additional random forest algorithm was performed to test for improvements. To preface, each attempt was conducted utilizing an 80/20 training and test split, which meant that the model was trained on 80% of the data before performing the test on the remaining 20%. Additionally, each test was conducted by setting a random state equal to 42. This figure allows the algorithm to be reproducible by designating the randomness level.

For the data about the year 2021, the random forest model generated a classification report to summarize the evaluation metrics. Out of 2428 predictions, 2255 were classified as true negatives, 74 as true positives, 92 as false positives, and seven as false negatives. Furthermore, the discrepancy between the class counts becomes evident when looking at the evaluation metrics for each class. On the non-fraud side, precision and recall were caricatured to be 1.0 and 0.96, respectively. On the other hand, for the fraud class, precision and recall drop heavily down to 0.43 and 0.92, respectively. These figures indicate an excellent model for classifying non-fraud cases, as seen by the high TN and accuracy scores. However, this model underperforms when predicting fraudulent claims, as explained by the low TP and fraud precision scores. These results are aligned with the expectations given the unbalanced nature of the outlier column. Furthermore, the support for non-fraud and fraud classes shows that 2219 entries belonged to the non-fraud class while only 209 belonged to fraud. This imbalance is also studied through decision tree analysis, which shows a significant class imbalance, especially when travelling down the leaves and nodes in each decision tree. Additionally, looking at the first tree produced in Python, it appears that the two features, “Avg_Spnd_Per_Dsg_Unt_Wghtd_2021” and “Tot_Benes_2021,” show the highest dominance for class 1 (fraud), indicated claims. Moreover, with the percentage dominance of 78%

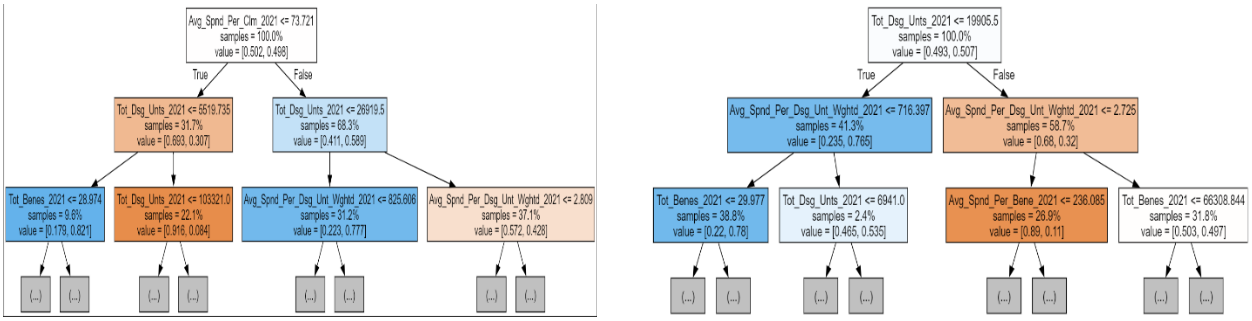
and 76.5% for both columns, the conclusion is that these two columns are essential features for explaining fraud in the medicare drug spending dataset. When looking at the second decision tree produced, “Tot_Clms_2021” is the feature with the highest Class (1) dominance, similar to the previously discussed features. Finally, when running a feature importance graph in Python, total claims, average spending per dosage unit, and average spending per beneficiary are essential in detecting fraud.

Table 1.6 2021 Classification Report (Non-SMOTE adjusted)



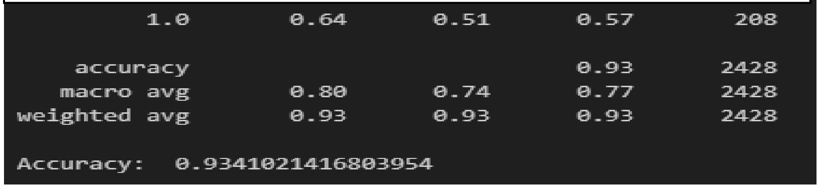
```
              precision    recall  f1-score   support

         0.0       1.00      0.96      0.98      2350
         1.0       0.43      0.92      0.59        78

    accuracy                           0.96      2428
   macro avg       0.72      0.94      0.78      2428
weighted avg       0.98      0.96      0.97      2428
```

Figure 1.7 Decision Tree Analysis 2021 Data (Non-SMOTE Adjusted)



Since the initial classification report seemed to underperform in the class 1 subsection, another attempt was run; this time, SMOTE was introduced to provide a balance to the fraud classes. Its implementation proved a class balance as non-fraud, and fraud had 9063 instances in the y sampled data. After re-running the random first model with these adjustments, a classification report was generated to view the improvements/deficiencies in the new approach. First, the accuracy score of 93.4% decreased as the model now has a more balanced distribution of correct classes to predict. In addition, for class 0 (non-fraud), the precision and recall scores remained significant at 0.95 and 0.97, respectively. However, for the previously dominated class (fraud), the synthetic balance significantly improved the recall score from 0.43 to 0.65. After SMOTE, the model is better at distinguishing the minority class (positive class) from the majority class because the class boundaries become more well-defined. Therefore, the number of false positives (cases incorrectly identified as fraud) decreases, improving precision. On the recall side, SMOTE focuses on generating synthetic data for the minority class compared to the majority. Since this does not add new features, the model may become slightly less sensitive to hard-to-classify positive instances. Furthermore, this can lead to a jump in the false negative (cases incorrectly identified as non-fraud), which drops the overall recall score.

Table 1.8 2021 Classification Report (SMOTE Adjusted)



```
         1.0       0.64      0.51      0.57       208

    accuracy                           0.93      2428
   macro avg       0.80      0.74      0.77      2428
weighted avg       0.93      0.93      0.93      2428

Accuracy:  0.9341021416803954
```

Figure 1.9 Decision Tree Analysis 2021 Data (SMOTE Adjusted)

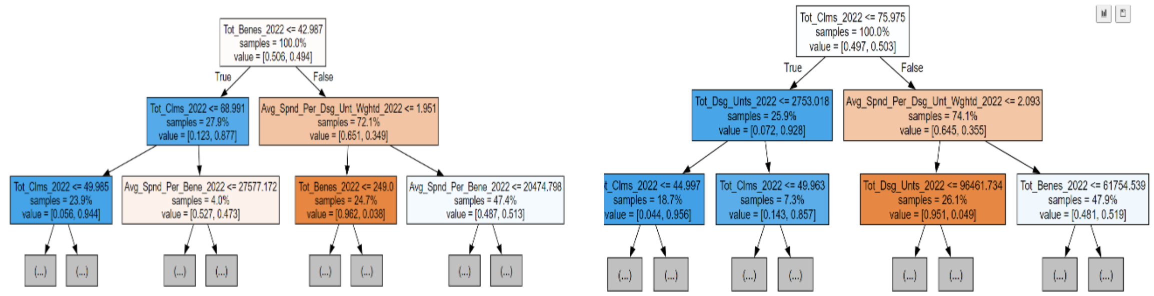
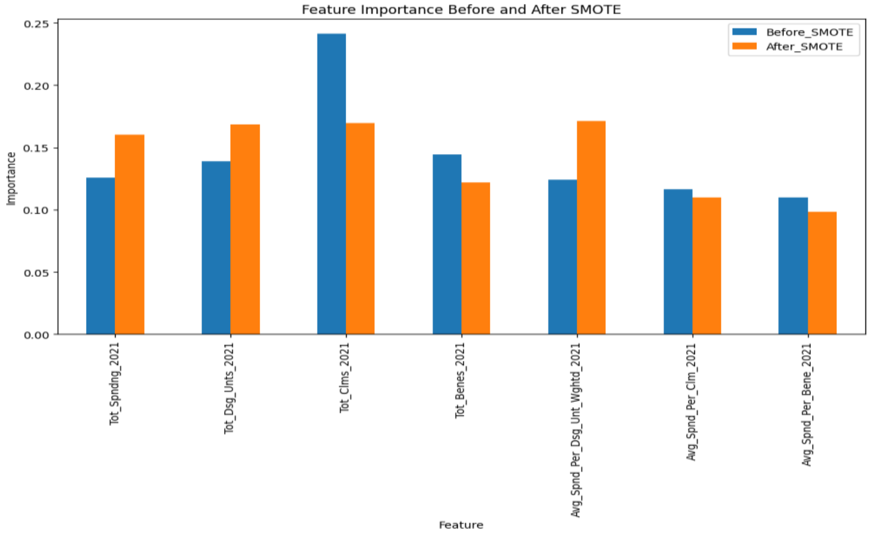
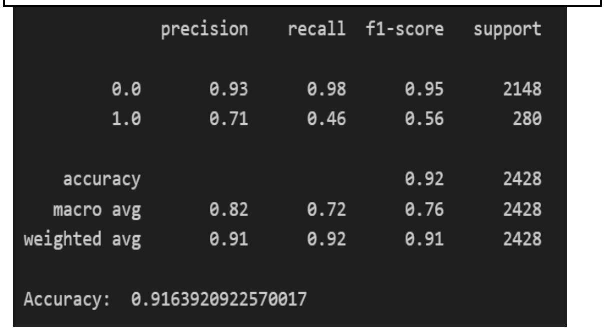


Figure 2.0 Feature Importance (2021 Data)



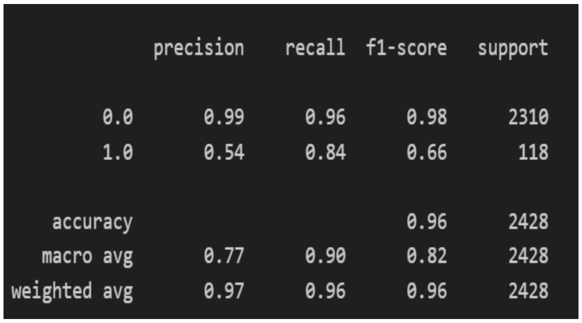
Since the data included columns for 2021 and 2022, it was imperative to mirror the previous models and analyze the 2022 data to get a more contemporary view. Similar metrics can be seen in 2022. Furthermore, the model begins favouring the precision/recall of the dominant (non-fraud) group with scores of 0.99 and 0.96, respectively. On the other hand, with a baseline low instance count, the fraud class attains a precision of 0.56 and a recall of 0.86. To improve the precision of the fraud class, the SMOTE is once again carried out, and this time, it produces a much more significant precision increase of 0.71. On the other hand, the recall did take a resounding negative hit, dropping it to 0.47. like before, the precision score increases compared to the drop in false positives, and the recall takes a hit as the false damage rises. When checking for vital features through decision tree analysis, the 2022 essential features are total claims, total beneficiaries, and weighted average spending per dosage unit. Overall, for both years, it is conclusive that a SMOTE analysis can benefit the random first model in that it offers more leeway and training for the model to accurately predict fraud and non-fraud instances within the Medicare prescription claims dataset.

Table 2.1 2022 Classification Report (Non-SMOTE Adjusted)



```
              precision    recall  f1-score   support

         0.0       0.93      0.98      0.95      2148
         1.0       0.71      0.46      0.56       280

    accuracy                           0.92      2428
   macro avg       0.82      0.72      0.76      2428
weighted avg       0.91      0.92      0.91      2428

Accuracy:  0.9163920922570017
```

Table 2.2 2022 Classification Report (SMOTE Adjusted)



```
              precision    recall  f1-score   support

         0.0       0.99      0.96      0.98      2310
         1.0       0.54      0.84      0.66       118

    accuracy                           0.96      2428
   macro avg       0.77      0.90      0.82      2428
weighted avg       0.97      0.96      0.96      2428
```

Figure 2.3 Decision Tree Analysis 2022 Data (Non-SMOTE Adjusted)

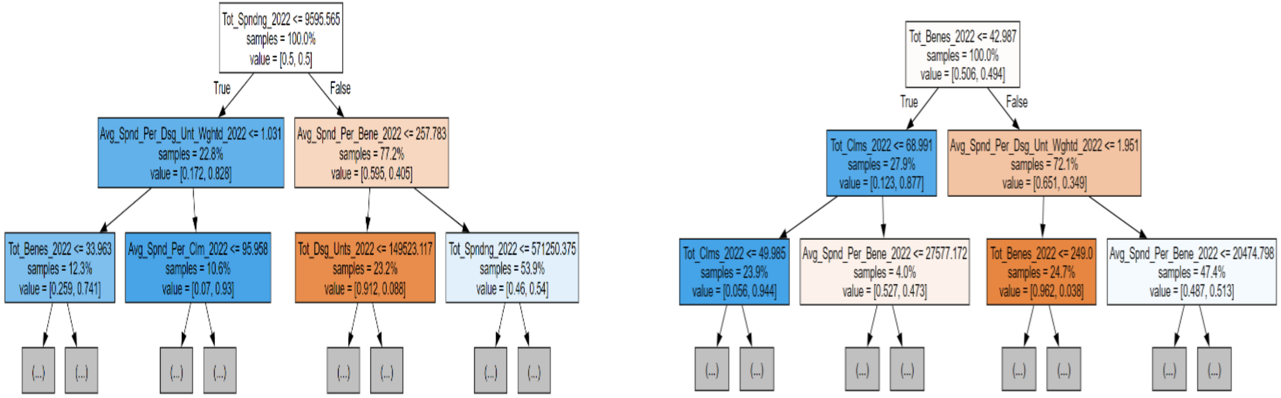
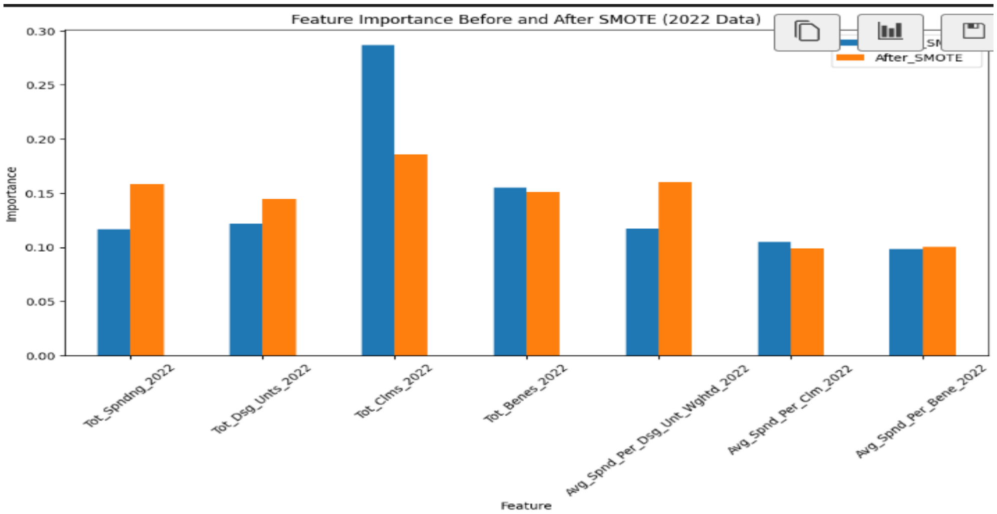


Figure 2.3 Feature Importance (2022 Data)



In addition to a random forest model capable of identifying fraud-given features, the research questions aimed to investigate more specific insights into the Medicare drug spending data set. Moreover, the implications of particular drug brands and their claim amounts can have a vital impact on the healthcare insurance industry. When looking at the fraud outlier counts of the top drugs in the United States, it emerges that Caspofungin Acetate, Modafinil, and Minocycline HCI ER are the top drugs with the highest fraud counts and, in turn, fraud rates. In terms of usage, Caspofungin is a medication used to treat several fungal infections in the organs, Modafinil is a widely prescribed sleep drug mainly given to those who suffer from narcolepsy (Rath 2024), and Minocycline HCI ER is a medication used to treat moderate to severe acne in people 12 years or older. All three drugs have general use cases that can appeal to the public and are popularly involved in fraud prescription claims. However, since the drug usage for the top 3 drugs is inconsistent, it was clear that additional analysis and insight were needed to classify common fraud-related prescriptions. With that, the original Medicaid dataset was aggregated to find the drugs with the highest claims and spending in dollar value. Results show that Eliquis (Apixaban), the standard treatment for blood clots, Trilucity (dulaglutide), a prescription for type 2 diabetes, and Revlimid (Lenalidomide), which is used to treat myeloma, score the highest in claim amounts and overall spending value in 2022 (Rath 2024).

2.4 Fraud Rates for top Drugs

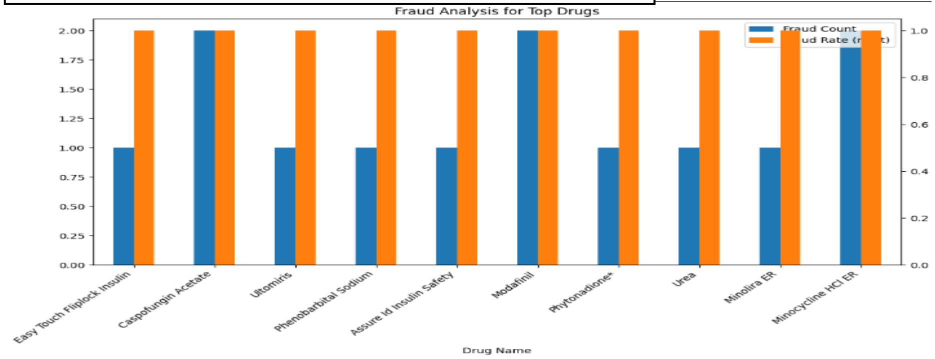


Figure 2.5 Total Spending vs. Total Claims

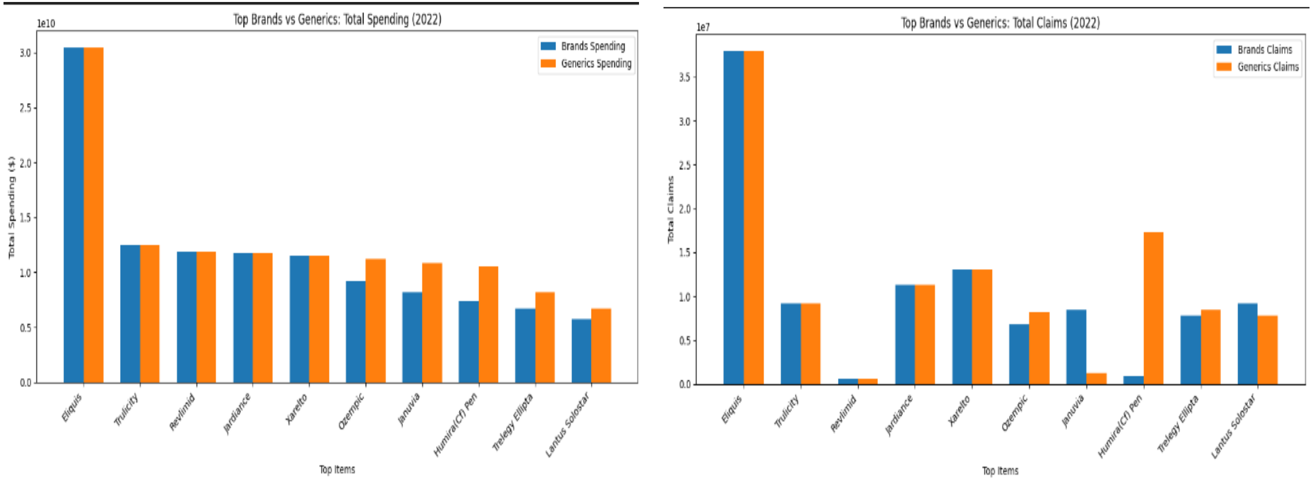
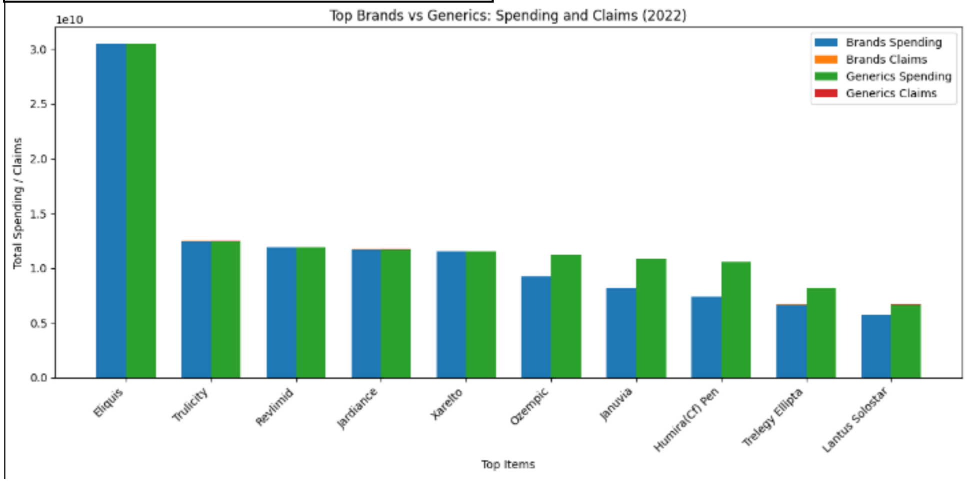


Figure 2.6



Discussion/Conclusion:

The analysis demonstrates that the random forest model effectively identified outliers in the Medicare Part D dataset. Key features such as total claims, average spending per dosage unit, and total beneficiaries were significant in determining fraudulent claims. While the baseline model provided substantial accuracy for the non-fraud class, using SMOTE improved the model's ability to detect minority (fraudulent claims, addressing the class imbalance, which enhanced the recall score. The main challenges with fraud detection were Class imbalance, feature complexity, and results interpretation. The attempts to improve detection will heavily depend on the slight improvement in those setbacks. In terms of drug types, certain drugs (e.g., Caspofungin Acetate, Modafinil, and Minocycline HCI ER) exhibited higher fraud rates, suggesting targeted misuse. In addition, Providers associated with these drugs showed varying prescribing patterns, indicating the potential need for deeper regulatory scrutiny and provider audits.

This research successfully applied machine learning techniques, specifically a random forest model coupled with SMOTE, to uncover fraudulent claims within the Medicare Part D dataset, identifying key predictors such as total claims and average spending per dosage unit. The study highlights high-risk drugs and providers and provides a roadmap for targeted fraud mitigation strategies. It emphasizes the critical role of data-driven approaches in improving healthcare policy efficiency. Policymakers, healthcare organizations, and pharmaceutical companies are urged to collaborate on implementing robust fraud detection systems, leveraging machine learning, data transparency, and enhanced audit protocols to safeguard Medicare's integrity. Beyond financial savings, addressing fraud ensures equitable healthcare service distribution and serves as a blueprint for global healthcare systems, promoting accountability and efficiency.

References

Aral, K. D., Guvenir, H. A., Sabuncuoglu, I., & Akar, A. R. (2011, November 15). *A prescription fraud detection model*. Computer Methods and Programs in Biomedicine. https://www.sciencedirect.com/science/article/abs/pii/S016926071100232X

Bayerstadler, A., Dijk, L. V., & Winter, F. (2016, September 30). *Bayesian multinomial latent variable modelling for fraud and abuse detection in health insurance*. Insurance: Mathematics and Economics. https://www.sciencedirect.com/science/article/abs/pii/S0167668715302845

Centers for Medicare & Medicaid Services Data. (2024, March 1). https://data.cms.gov/resources/medicare-part-d-spending-by-drug-methodology

Centers for Medicare & Medicaid Services Data. (n.d.). https://data.cms.gov/summary-statistics-on-use-and-payments/medicare-medicaid-spending-by-drug/medicare-part-d-spending-by-drug

FBI. (2016, June 1). *Health Care Fraud*. FBI. https://www.fbi.gov/investigate/white-collar-crime/health-care-fraud

Haque, M. E., & Tozal, M. E. (n.d.). *IEEE.pdf*. Identifying Health Insurance Claim Frauds Using Mixture of Clinical Concepts. https://www.iitrpr.ac.in/library/pdf/IEEE.pdf

Iyengar, V. S., Hermiz, K. B., & Natarajan, R. (2013, July 3). *Computer-aided auditing of prescription drug claims*. Health care management science. https://pubmed.ncbi.nlm.nih.gov/23821344/

Johnson, M. E., & Nagarur, N. (2015, January 20). *Multi-stage methodology to detect health insurance claim fraud - health care management science*. SpringerLink. https://link.springer.com/article/10.1007/s10729-015-9317-3

Joudaki, H., Rashidian, A., Minaei-Bidgoli, B., Mahmoodi, M., Geraili, B., Nasiri, M., & Arab, M. (2015, November 10). *Improving fraud and abuse detection in general physician claims: A Data Mining Study*. International journal of health policy and management. https://pubmed.ncbi.nlm.nih.gov/26927587/

Kumaraswamy, N., Markey, M. K., Barner, J. C., & Rascati, K. (2022, August 8). *Feature engineering to detect fraud using healthcare claims data*. Expert Systems with Applications. https://www.sciencedirect.com/science/article/abs/pii/S0957417422015330

Medicaid Spending by Drug. Centers for Medicare & Medicaid Services Data. (n.d.). https://data.cms.gov/summary-statistics-on-use-and-payments/medicare-medicaid-spending-by-drug/medicaid-spending-by-drug

Rath, K. (2024, July 21). *Eliquis (apixaban): Uses, side effects, interactions, pictures, warnings & dosing*. WebMD. https://www.webmd.com/drugs/2/drug-163073/eliquis-oral/details

Settipalli, L., & Gangadharan, G. R. (2022, November 19). *WMTDBC: An unsupervised multivariate analysis model for fraud detection in health insurance claims*. Expert Systems with Applications. https://www.sciencedirect.com/science/article/abs/pii/S0957417422022771

Shin, H., Park, H., Lee, J., & Jhee, W. C. (2012, January 24). *A scoring model to detect abusive billing patterns in health insurance claims*. Expert Systems with Applications. https://www.sciencedirect.com/science/article/abs/pii/S0957417412001236#:~:text=We%20propose%20a%20scoring%20model%20that%20detects%20outpatient,categorize%20the%20problematic%20providers%20with%20similar%20utilization%20patterns.

TI; G. N. H. E., Oh, H., Rehmet, E., & Shireman, T. I. (2024, July 10). *Descriptive trends in Medicaid antipsychotic prescription claims and expenditures, 2016 - 2021*. The journal of behavioural health services & research. https://pubmed.ncbi.nlm.nih.gov/38987413/

Zafari, B., & Ekin, T. (2018, December 16). *Topic modelling for medical prescription fraud and abuse detection*. OUP Academic. https://academic.oup.com/jrsssc/article/68/3/751/7058372