Unveiling Temporal Disparities in Medicare DME: Machine Learning with Time-Series Cross-Validation for Cost Efficiency and Resource Allocation

Submitted by: Rozani Jeganathan [501282785]

Supervisor: [Supervisor's Name]

Date of Submission: 22 Sep 2024

Institution/Department: The G. Raymond Chang School of Continuing Education

Toronto Metropolitan University

Abstract

Fraudulent activities within the U.S. healthcare system, particularly in Medicare, result in substantial financial losses. In 2023, Medicare disbursed an estimated $31.2 billion in improper payments, representing 7.38% of total program expenditures. These improper payments, which include fraud, abuse, and billing errors, pose significant challenges to the system's efficiency and sustainability [1,2].

To mitigate these financial losses, machine learning techniques are applied to detect fraud by identifying unusual patterns in supplier charges and service volumes. This study analyzes nearly 10 years of Medicare Durable Medical Equipment, Prosthetics, Orthotics, and Supplies (DMEPOS) data [3] to uncover inefficiencies and identify underserved regions. The extended time frame allows for a comprehensive analysis of long-term trends and geographic disparities in billing practices. While numerous studies have focused on Medicare datasets such as Part B, Part D, and other DMEPOS subsets, this research uniquely focuses on a dataset with detailed regional information, offering valuable insights into geographic variations in service delivery and potential inefficiencies.

By focusing exclusively on DMEPOS data and utilizing both supervised and unsupervised machine learning models, this study aims to detect anomalies, analyze correlations between supplier service volumes and charges, and address imbalances in Medicare resource allocation. Unlike previous studies that often rely on shorter timeframes or combine multiple datasets, our research provides a deeper exploration of long-term trends in DMEPOS services across various geographic regions. This approach offers critical insights into how temporal patterns in supplier

behavior can reveal inefficiencies and inform Medicare on optimizing costs, ensuring affordability, and improving accessibility for underserved communities.

This decade-long dataset, with its regional focus, provides a robust framework for understanding DMEPOS services and supplier behaviors across different regions, particularly underserved areas. The study highlights geographic disparities in billing practices and analyzes temporal trends in service distribution, helping to optimize Medicare resource allocation and ensure equitable access to DMEPOS products and services. Compared to prior research, which focuses on shorter periods or multiple Medicare datasets, this study offers region-specific insights into supplier behaviors, recurring fraud patterns, and potential inefficiencies within the DMEPOS sector.

By leveraging machine learning and time-series analysis, our study contributes to more accurate detection of fraudulent activities, inefficiencies, and underserved regions within the Medicare system. These findings are expected to inform policy decisions and improve Medicare's cost efficiency, resource allocation, and regional equity in the delivery of DMEPOS service.

The project addresses the following research questions:

1. How do aggregated supplier behavior patterns (in terms of service volumes and Medicare payments) change over time, and which time periods exhibit signs of inefficiencies or potential fraud?
2. Are certain regions consistently associated with higher supplier service volumes and Medicare payments, and are higher rates charged in 'underserved' areas, particularly in specific time periods?

3. How have supplier charges varied by region over time, and how can these temporal trends help optimize costs?
4. How do correlations between service volumes and supplier charges evolve over time, and what temporal patterns suggest inefficiencies?
5. What are the temporal trends in DME service distribution across geographic regions, and how can these trends identify underserved regions over time?

By using Python, R, and SAS for data analysis, and employing visualization techniques with Matplotlib, Seaborn, Tableau, and Power BI, this project demonstrates the application of advanced machine learning techniques to inform Medicare policy development. Tools such as Jupyter, Colab Notebooks are used for iterative development, combining code, text, and visualizations. Additionally, Pandas and NumPy are essential for data manipulation and preprocessing, while Scikit-learn is employed for implementing supervised and unsupervised machine learning models. The findings will provide actionable insights for improving resource allocation, optimizing healthcare access, and identifying regions or suppliers exhibiting inefficiencies or potential fraud.

Keywords: *Medicare DME, Classification, Regression, Anomaly Detection, Cost Efficiency, Machine Learning, Time-Series cross-validation, Resource Allocation, Fraud Detection.*

Literature Review

Introduction

Fraudulent activities in healthcare, particularly within Medicare, represent a significant challenge, leading to financial losses and inefficiencies in resource allocation. Over the past decade, there has been a growing interest in employing machine learning and time-series analysis to detect anomalies in Medicare data, optimize cost efficiency, and improve healthcare access. Researchers have primarily focused on various Medicare datasets, including Part B, Part D, and Durable Medical Equipment, Prosthetics, Orthotics, and Supplies (DMEPOS), to analyze supplier behavior and service volumes. While regional variations are often implied, few studies have explicitly explored geographic disparities, leaving a gap in understanding how location impacts inefficiencies and fraudulent activities within the system.

A variety of methodologies have been proposed, with studies leveraging supervised and unsupervised machine learning models, such as Random Forest, Gradient Boosting, Logistic Regression, Isolation Forest, and Local Outlier Factor (LOF), to identify fraudulent patterns and inefficiencies. Many researchers assume that fraud can be detected through the identification of anomalies in provider billing and service patterns, often utilizing combined datasets to enhance detection accuracy. In addition, time-series forecasting models like ARIMA and SARIMA have been used to capture temporal trends in Medicare fund flows and provider behavior.

Despite the growing consensus on the effectiveness of machine learning and time-series models, conflicting views exist regarding the most effective methods. While traditional models such as

Random Forest and Logistic Regression are widely used, some researchers advocate for more advanced approaches like deep learning and cost-sensitive learning to improve fraud detection and handle imbalances in data. This literature review examines key findings from recent studies, highlighting the assumptions, methodologies, and areas where conflicting theories or methodologies emerge, offering a comprehensive view of current research in Medicare fraud detection and resource optimization.

Articles

In the peer-reviewed article [4], “Big Data Fraud Detection Using Multiple Medicare Data Sources” (Herland, M., Khoshgoftaar, T. M., & Bauder, R. A. 2018), the authors explore the use of machine learning techniques to detect fraud within the U.S. Medicare system by combining datasets from three key Medicare sources: Part B (Physician and Supplier), Part D (Prescription Drug), and DMEPOS (Durable Medical Equipment, Prosthetics, Orthotics, and Supplies). Their objective was to determine the most effective approach to detect fraudulent activity among healthcare providers. The researchers combined Medicare claims data from 2012 to 2016 with fraud labels from the List of Excluded Individuals and Entities (LEIE) to classify providers as fraudulent or non-fraudulent. Independent variables included billing patterns, service volumes, HCPCS codes, provider types, and payment data, while the dependent variable was whether a provider appeared on the LEIE fraud list. The researchers utilized three machine learning models (Logistic Regression, Random Forest, and Gradient Boosted Trees) on both individual and combined datasets. They found that Logistic Regression performed best, particularly when applied to the combined dataset, yielding an AUC score of 0.816. The study emphasizes the

value of integrating multiple data sources for better fraud detection, as focusing on a single dataset may overlook fraudulent patterns in other Medicare areas. Integrating multiple datasets improved fraud detection performance compared to using individual datasets, highlighting the need for a comprehensive approach in identifying fraudulent provider behaviors. This approach is consistent with the objectives of our research, as we similarly seek to detect anomalous patterns in Medicare data, though our emphasis is on identifying inefficiencies and underserved regions. While Herland et al. employ machine learning models to identify fraudulent activity, our study applies similar techniques to analyze anomalies in supplier charges and service volumes, with the goal of uncovering inefficiencies within the system.

In the scholarly study [5], "Medicare Fraud Detection Using Machine Learning Methods" (Bauder, R. A., & Khoshgoftaar, T. M. 2017), the authors compare several machine learning approaches to detect fraud within the U.S. Medicare system. They examine supervised, unsupervised, and hybrid models applied to Medicare data from 2015, using fraud labels from the List of Excluded Individuals/Entities (LEIE) database. The study aimed to determine the most effective method to classify fraudulent providers by employing models like Gradient Boosted Machines (GBM), Random Forest (RF), Deep Neural Networks (DNN), and Naive Bayes for supervised learning, while using Autoencoders, Mahalanobis distance, k-Nearest Neighbors (kNN), and Local Outlier Factor (LOF) for unsupervised learning. A hybrid approach involving an autoencoder-pre trained neural network was also used. The authors focused on the severe class imbalance in the Medicare data, addressing it with oversampling and an 80-20 sampling technique. The primary metrics used were Balanced Accuracy (BACC), F-measure, G-measure, and Matthew's Correlation Coefficient (MCC), which were employed to evaluate the

models. Among the models tested, the 80-20 sampling method demonstrated the best performance, with supervised learners outperforming both unsupervised and hybrid models, particularly in detecting fraudulent providers. The study concluded that supervised models, especially Gradient Boosted Machines and Random Forests, were most effective in detecting fraud in specialized provider types. This work is closely related to my project, as both aim to detect anomalies in Medicare data using machine learning models. While the article focuses on fraud detection among providers, our project specifically targets anomalies in supplier charges and service volumes to identify inefficiencies and underserved regions. Both studies address the issue of data imbalance and employ supervised learning methods, such as Random Forest, which is also central to our approach for identifying unusual patterns in the data. Additionally, unsupervised learning methods like k-Nearest Neighbors (kNN) are relevant to both analyses for detecting anomalous behaviors.

In this academic research [6], "Identifying Medicare Provider Fraud with Unsupervised Machine Learning" (Bauder, R. A., da Rosa, R. C., & Khoshgoftaar, T. M. 2018), the authors explore various unsupervised machine learning methods to detect fraud in Medicare claims data. The study focuses on Medicare Part B data from 2012 to 2015 and uses the List of Excluded Individuals/Entities (LEIE) database for validation. The authors assess five unsupervised machine learning algorithms: Isolation Forest, Unsupervised Random Forest, Local Outlier Factor (LOF), k-Nearest Neighbors (KNN), and autoencoders. Each method was tested for its ability to detect outliers, which could indicate fraudulent providers. The study utilized provider-level features such as billing amounts, number of claims, types of services rendered, and service frequencies. The dataset was unlabeled, and the goal was anomaly detection to identify potential

fraud without direct fraud labels for training. The study's results showed that the Local Outlier Factor (LOF) outperformed the other methods, achieving the highest Area Under the ROC Curve (AUC) score of 0.63. In contrast, Isolation Forest and autoencoders had the lowest AUC scores, performing close to random guessing. The study highlights the challenge of working with highly imbalanced datasets, as fraudulent instances made up only 0.04% of the data. LOF was particularly effective in detecting fraudulent providers, though other methods like Unsupervised Random Forest and KNN also showed promise with different configurations. The article concludes that while unsupervised methods can help detect fraud when labeled data is unavailable, the results indicated that these methods are still less effective compared to supervised learning. The authors suggest further research, including testing on larger datasets and improving the balance of fraud and non-fraud instances, could enhance detection performance. This work relates to our project as both focus on detecting anomalies in healthcare data using unsupervised machine learning models. While the article targets fraudulent providers in Medicare Part B, our project identifies inefficiencies and underserved regions by analyzing supplier charges and service volumes. Both studies face challenges with imbalanced datasets and use models like Isolation Forest and Local Outlier Factor (LOF) to detect outliers. Additionally, both focus on provider-level features and detecting unusual patterns in large datasets without labeled data.

In this reviewed study [7], "Cost-Sensitive Learning for Medical Insurance Fraud Detection with Temporal Information" (Shi, H., Tayebi, M. A., Pei, J., & Cao, J. 2023), the authors propose a novel framework to detect fraudulent activities in Medicare claims by utilizing temporal information and a cost-sensitive learning approach. The study focuses on Medicare Fee-For-

Service (FFS) claims data, using data from previous years to form temporal trajectories of key covariates such as claim counts and costs. The authors apply Functional Principal Component Analysis (FPCA) to extract features from these temporal covariates, enabling the analysis of patterns over time. A key innovation of this study is the incorporation of cost-sensitive learning to account for the asymmetric financial losses associated with false negatives (failing to detect fraud) versus false positives (wrongly identifying legitimate claims as fraud). This framework allows the detection method to balance the trade-off between the cost of investigating potential fraud and the actual fraud losses. Class imbalance, a common issue in fraud detection, is tackled through a random under sampling method, where multiple class ratios are evaluated to improve the classifier's performance. The authors tested four machine learning algorithms—Logistic Regression, Random Forest, Gradient Boosting Machine (GBM), and Neural Networks—under different under sampling schemes. The Random Forest and GBM models, particularly under a 5% fraud ratio (RUS-5), demonstrated the highest cost savings and predictive accuracy, with Random Forest achieving steady performance across multiple metrics. The results indicate that cost-sensitive learning combined with temporal trajectory analysis can significantly reduce fraud detection costs, achieving up to 55% savings compared to non-cost-sensitive approaches. This study closely relates to our project as both emphasize temporal analysis and cost-sensitive learning to detect anomalies in healthcare data. While the article focuses on fraudulent activities in Medicare claims, our project targets inefficiencies and underserved regions by analyzing supplier charges and service volumes over time. Both face challenges like class imbalance and leverage temporal data to uncover patterns indicating unusual behavior. Additionally, the use of Random Forest and Gradient Boosting Machine (GBM) parallels the models we are considering,

with cost-sensitive learning offering valuable insights for optimizing resource allocation in our project.

In this peer revied article [8], “Time Series Analysis and Forecasting of Monkeypox Disease Using ARIMA and SARIMA Models” (Pramanik, A., Sultana, S., & Rahman, M. S., 2022), the authors propose and evaluate the use of ARIMA and SARIMA models to forecast the spread of Monkeypox disease. The study is based on global Monkeypox case data from January 2022 to September 2022. The key aim of the study was to predict future trends in Monkeypox cases and to provide insight into managing public health strategies for the outbreak. The models applied were Auto-Regressive Integrated Moving Average (ARIMA) and Seasonal ARIMA (SARIMA). The research involved preprocessing data to ensure it was suitable for time-series analysis, including making the data stationary by removing seasonal components, and fitting the ARIMA and SARIMA models. The study evaluated the models based on performance metrics such as Root Mean Square Error (RMSE), and found that the SARIMA model outperformed ARIMA, providing better forecasting accuracy with an RMSE of 3.1180 compared to 3.6818 for ARIMA. This study highlights the importance of using time-series models to capture both seasonality and trends when predicting the spread of diseases, contributing to effective public health planning and resource allocation. The application of these time-series techniques can be relevant to other healthcare forecasting problems, such as predicting anomalies in healthcare datasets or service demands over time. This study relates to our project as it applies time-series analysis to forecast trends in healthcare data, aligning with our goal of analyzing Medicare supplier charges and service volumes over time. While the article predicts the spread of Monkeypox using ARIMA and SARIMA models, our project similarly focuses on identifying patterns and anomalies in

Medicare data, particularly supplier charges and inefficiencies in underserved regions. Both projects deal with seasonal and temporal trends, making SARIMA especially relevant for analyzing supplier charges across regions. The use of metrics like RMSE to evaluate forecasting accuracy also informs our assessment of time-series models in predicting supplier behavior and optimizing costs.

In the research article [9], "Time Series Forecasting of Medicare Fund Expenditures Based on Historical Data—Taking Dalian as an Example" (Qu, G., Cui, S., & Tang, J. (2014)) utilizes time series forecasting to predict Medicare fund expenditures based on historical data. The study focuses on the imbalance between Medicare fund revenues and expenditures, using time-series methods to anticipate trends and inform policy adjustments. The application of time series analysis aligns with our project as we also aim to analyze temporal trends in Medicare supplier charges and service volumes over time. While this article addresses Medicare fund forecasting in China, the methodologies, including forecasting and managing health-related expenditures, provide insights into managing temporal data in healthcare, which can be applied to identifying inefficiencies and predicting future supplier behaviors in our project.

Conclusion

This literature review reveals significant findings regarding the methodologies and assumptions surrounding fraud detection in Medicare. Most researchers operate under the assumption that fraudulent activities can be identified through anomalies in provider billing and service patterns. While various machine learning models, including Random Forest, Isolation Forest, and

SARIMA, are widely discussed, conflicting views arise regarding the effectiveness of traditional versus advanced techniques.

Many studies focus primarily on fraud detection rather than an in-depth analysis of supplier behavior and service volumes, highlighting a notable gap in the literature. Specifically, while aspects of supplier behavior and service volumes are mentioned, none of the reviewed articles provide a comprehensive examination of these factors, especially in relation to geographic disparities. This presents a compelling case for further investigation into how location impacts inefficiencies within the Medicare system.

Our research aims to fill this gap by leveraging the identified methodologies and focusing on supplier behavior and service volumes to uncover inefficiencies and underserved regions in Medicare datasets. By applying machine learning models and time-series analysis, we aim to contribute valuable insights into the long-term trends affecting Medicare and inform strategies for optimizing resource allocation. The findings of this review underscore the importance of addressing these overlooked areas to enhance the understanding of Medicare fraud and inefficiencies.

Methodology

Based on insights from prior research, a preliminary overall methodology has been developed to effectively address the research questions presented in the abstract section. Figure 1 below outlines the steps involved in this methodology

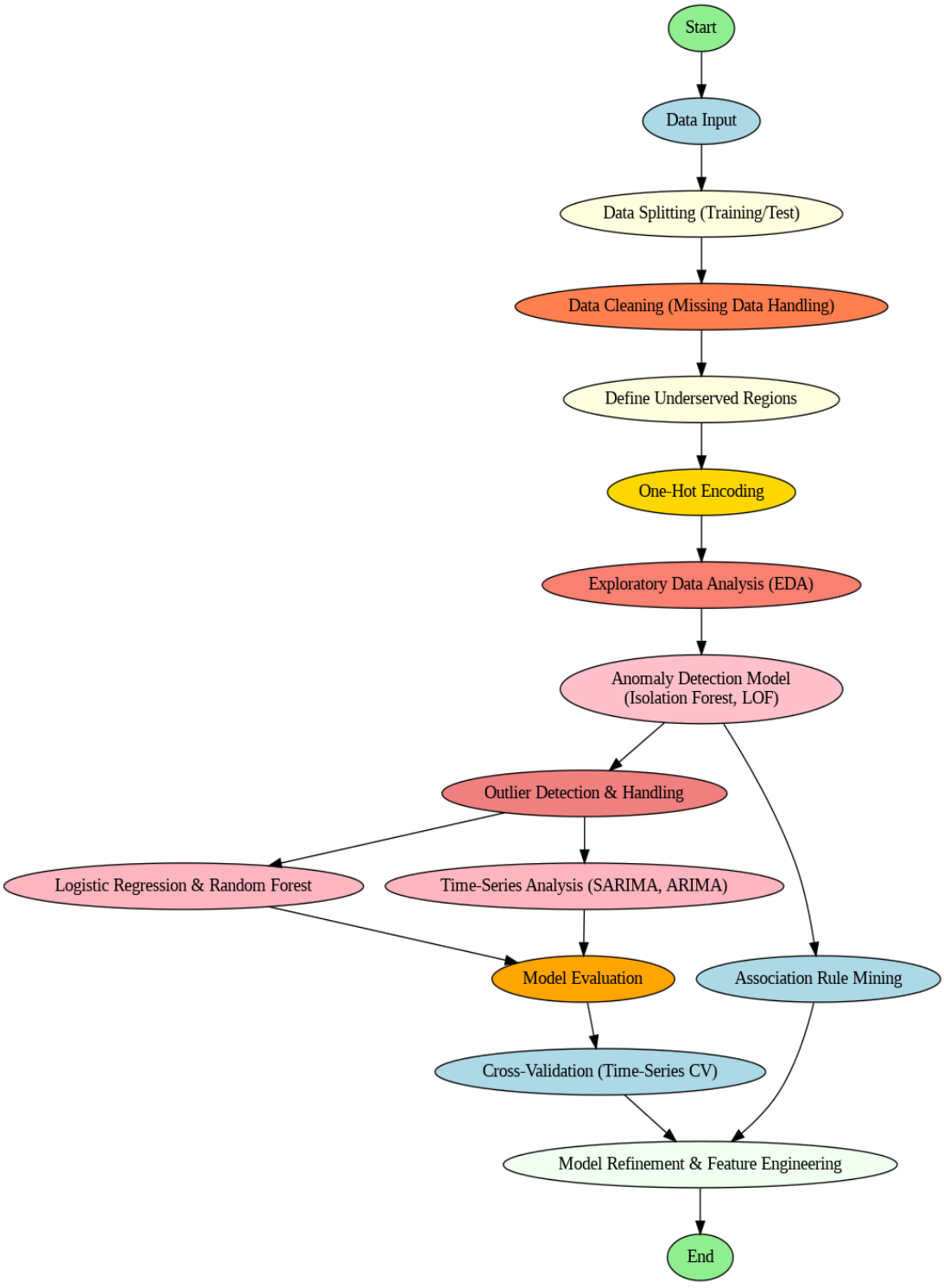

Figure 1 - Tentative overall methodology graph

Data

Data Understanding

This section describes the Medicare Durable Medical Equipment, Devices & Supplies - by Geography and Service dataset used in our project. This DMEPOS dataset provides detailed information on durable medical equipment, prosthetics, orthotics, and supplies that Medicare Part B patients can purchase or rent from suppliers, as referred by physicians. It contains records of services provided to Medicare Part B (fee-for-service) beneficiaries, ordered by healthcare professionals, and aggregated by geography and service type. The dataset spans nearly a decade, from 2014 to 2022, with 18 variables and contains no duplicate values.

Data Splitting

Training set

In this project, the Training Set includes data from 2014 to 2020, which allows the model to understand historical supplier behavior, service volumes, and Medicare payments before being validated on future unseen data.

Test Set

In this project, the Test Set includes data from 2021 to 2022, providing a way to assess how well the model predicts and performs on recent data, particularly in detecting anomalies, inefficiencies, and underserved regions.

The dataset will be divided into Training, Validation, and Test sets based on the year of data as outlined below:

Data year	Training Period	Blocked Cross-validation Period	Test Period
2014 - 2020	2014 - 2015	2016	_
	2014 to 2017	2018	_
	2014 to 2018	2019	_
	2014 to 2019	2020	_
2021 - 2022	_	_	2021 - 2022

Table 1 - Data Segmentation for Model Training and Testing Periods

Blocked Cross-validation is used in this project due to the temporal nature of the dataset, where data from different years may exhibit distinct patterns over time. Unlike standard cross-validation, which randomly splits data into training and validation sets, Blocked Cross-validation preserves the temporal order of the data, preventing data leakage from the future into the past. This method is particularly crucial when working with time-series data, as it ensures that future information does not influence the model during training.

Data Processing

The data underwent several processing steps, including cleaning, handling missing values, and applying one-hot encoding for categorical variables. So far, these steps have been applied only to the training set. The test set will undergo the necessary preprocessing after model validation.

Missing Value Handling

As part of the data preparation process, datasets spanning from 2014 to 2020 were merged, with missing values addressed systematically before combining the data. This ensured that critical regional information was retained, minimizing the impact of missing data.

Throughout the dataset, missing values were primarily observed in three columns: Referring Provider Geography Description, Referring Provider Geography Code, and Total Supplier Beneficiaries. All three columns exhibited similar patterns of missing data.

- For the Referring Provider Geography Description column, which had less than 0.02% missing data, the rows with missing values were removed.
- For missing values in the Referring Provider Geography Code, a placeholder string "National" was used to maintain the regional context. This adjustment was necessary as the national column lacks a regional code, ensuring the dataset remains complete and suitable for model training.
- For Total Supplier Beneficiaries, which had 10-12% missing data, the median was used as the replacement value. Given the left-skewed distribution, the median provides a more robust and accurate measure than the mean.

This consistent approach across all columns ensured that critical data integrity was preserved while preparing the dataset for further analysis.

Initially, two datasets were created, one where missing value were removed and another where missing values for Total Supplier Beneficiaries were replaced with the median. The dataset with the median replacement is currently planned for model training, as the dataset with removed NA values would lead to the loss of important regional information, as highlighted during the EDA

analysis. However, KNN Imputation may be considered later, depending on validation results, to further refine the dataset and improve model performance.

Combined Data

The combined dataset includes data from 2014 to 2020, with 295,052 observations and 19 variables, including a discrete year variable to represent each year.

Identifying Underserved Regions

In our project, underserved regions are identified by a combination of high supplier charges and low service volumes. Specifically, regions where supplier charges fall within the 95th percentile or higher, and service volumes are in the 5th percentile or lower are flagged. This dual condition highlights areas with unusually high costs but limited access to services. The identification process utilizes three newly created variables: high_charges, which indicates whether average supplier submitted charges are in the 95th percentile or higher; low_service_volumes, which flags regions where total supplier services are in the 5th percentile or lower; and is_underserved, which marks regions that meet both conditions simultaneously. This approach indicates potential inefficiencies or disparities in service distribution.

Since the dataset does not contain National Provider Identifier (NPI) information, we are unable to directly map fraud labels from external sources such as the LEIE (List of Excluded Individuals/Entities). This limitation prevents the application of supervised learning techniques for fraud detection. As a result, our analysis focuses on unsupervised learning methods, utilizing

features such as service volumes, payments, and geographic data to detect potential anomalies or suspicious behavior without relying on labeled fraud data.

Summary statistic of Combined data

A summary of the descriptive statistics for the combined dataset, encompassing 22 features, including the added Year and Underserved columns (high_charges, low_service_volumes, is_underserved), was compiled and is presented in Table 2 & 3.

Variable	Mean	Median	Min	Max	Standard Deviation	Type
Rfrg_Prvdr_Geo_Lvl	-	-	-	-	-	Character
Rfrg_Prvdr_Geo_Cd	-	-	-	-	-	Character
Rfrg_Prvdr_Geo_Desc	-	-	-	-	-	Character
RBCS_Lvl	-	-	-	-	-	Character
RBCS_Id	-	-	-	-	-	Character
RBCS_Desc	-	-	-	-	-	Character
HCPCS_Cd	-	-	-	-	-	Character
HCPCS_Desc	-	-	-	-	-	Character
Suplr_Rentl_Ind	-	-	-	-	-	Character
Tot_Rfrg_Prvdrs	478.6	47	1	271965	4,182.21	Integer

Tot_Suplrs	94.38	17	1	49095	686.05	Integer
Tot_Suplr_Benes	1669	89	11	3495668	25,827.94	Integer
Tot_Suplr_Clms	4371	121	11	10498977	76,475.00	Integer
Tot_Suplr_Srvcs	83082	305	11	60884546	2,317,080.0	Integer
Avg_Suplr_Sbmtd_Chrg	541.91	92.21	0.01	70913.34	1,779.18	Numeric
Avg_Suplr_Mdcr_Alowd_Amt	340.29	48.73	0.01	39757.35	1,185.54	Numeric
Avg_Suplr_Mdcr_Pymt_Amt	263.38	36.761	0	31071.56	919.36	Numeric
Avg_Suplr_Mdcr_Stdzd_Amt	264.21	37.475	0	31019.03	918.39	Numeric
Year	2017	2017	2014	2020	2.00	Numeric

Table 2 – Summary statistic of Combined data

Variable	No of False	Percentage	No of True	Percentage	Type
high_charges	280299	95.0%	14753	5.0%	Logical
low_service_volume	277781	93.5%	17271	6.5%	Logical
is_underserved	292984	99.3%	2068	0.7%	Logical

Table 3 – Summary statistic of underserve variables

A bar plot of the binary categorical variable illustrates the distribution of its two categories, highlighting the counts of each category for easy comparison.

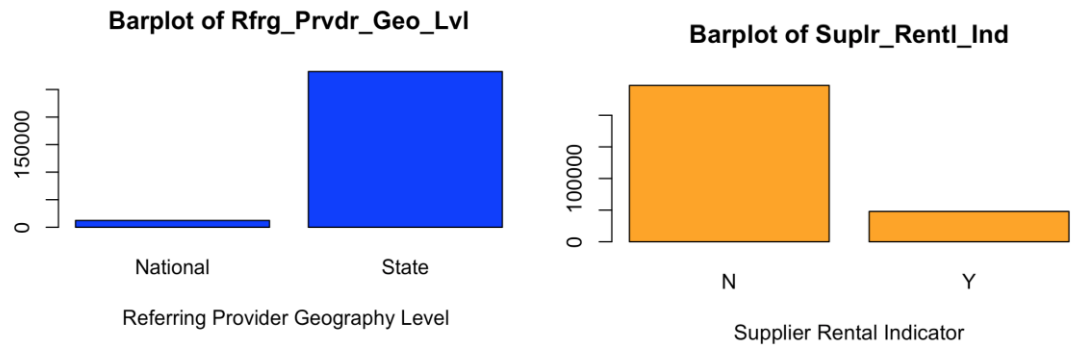

Figure 2 – Bar plot of Binary categorical variable

One-Hot Encoding

In this research, one-hot encoding was applied to transform categorical features into binary columns for model building. This process initially encompassed all categorical variables, resulting in a significant expansion of the dataset's dimensionality, with 3,834 columns. However, when applying PCA, the high dimensionality of the dataset caused an excessive computation time. To effectively manage this complexity and prepare for subsequent analysis, the number of categorical variables included in one-hot encoding was strategically reduced based on the unique values in each column. At least one column was selected from each group of related variables, ensuring that important connections among categories were preserved while reducing redundancy. In this way the number of columns reduced to 1810. Principal Component Analysis (PCA) was then applied to further reduce dimensionality while retaining the most

informative features, as discussed in the exploratory data analysis (EDA). This approach-maintained data integrity while uncovering key insights that informed the analysis.

Exploratory Data Analysis

Exploratory data analysis (EDA) was conducted to individual dataset to gain an initial understanding of the dataset, using GitHub Copilot. Important insights related to missing values and outliers were uncovered, highlighting potential inefficiencies that warrant further investigation.

Additionally, feature analysis was performed to combined data set using R Studio and the datamaid package, providing a comprehensive overview of the data's structure, quality, and key variables, setting the foundation for more detailed analyses.

All continuous variables exhibit right-skewness and do not follow a normal distribution, as visualized through histograms. Various transformations, including square root, cube, log, and scaling methods, were applied to reduce the skewness of the continuous variables. While these methods improved the distribution, the data remained non-normally distributed according to the Shapiro-Wilk test. This suggests that further non-parametric methods may be needed for analysis.

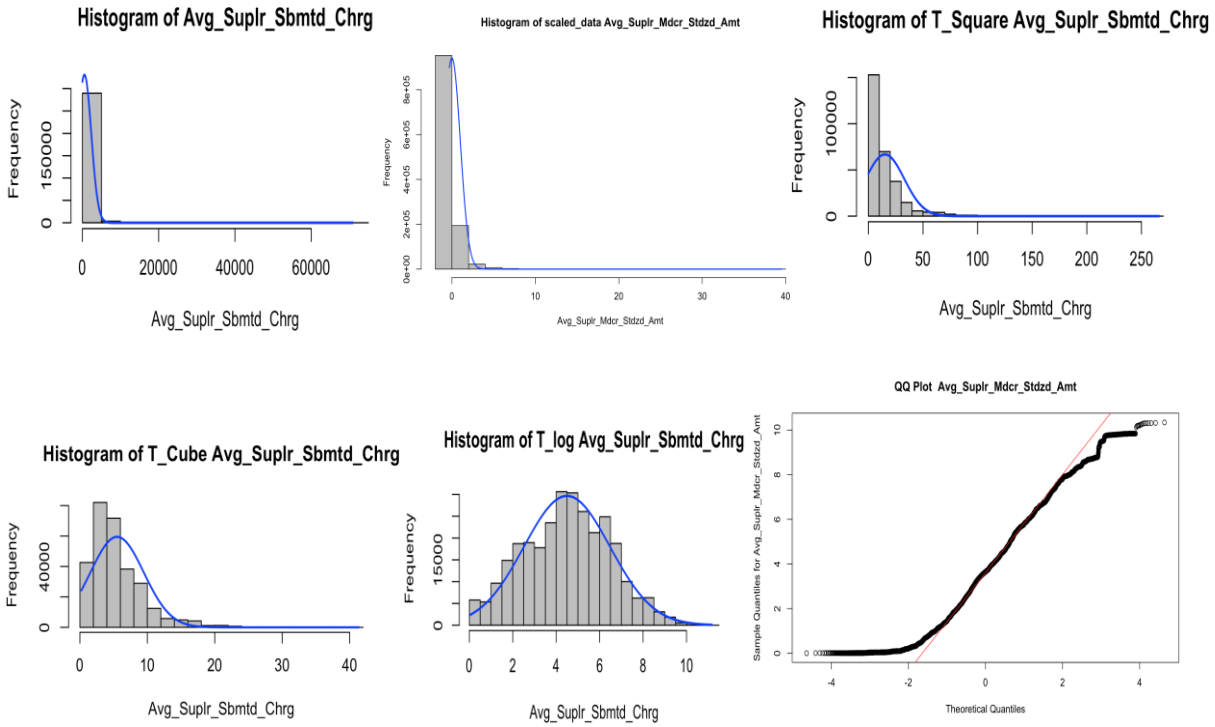

Figure 3 – Histograms and QQ Plot of Average Supplier Submitted Charge

Due to the right-skewness and non-normal distribution of most discrete variables, as confirmed by the Shapiro-Wilk test and visualized through histograms, applied Spearman's rank correlation for all the numerical columns to assess relationships between variables.

Correlation analysis

For the correlation analysis, the focus was placed on the numeric columns within the dataset to uncover meaningful relationships. Spearman's rank correlation was employed to analyze the relationships between variables.

Correlation Analysis of Original Numeric Columns

The high correlation among Tot_Rfrg_Prvdrs, Tot_Suplrs, and Tot_Suplr_Benes suggests a strong relationship between the number of referring providers and the total number of suppliers and beneficiaries. Conversely, the weak correlation with the Year variable indicates that these metrics are relatively stable over time within the dataset.

All the continues variable are highly correlated with one another, with correlation coefficients consistently above **0.95**. This strong correlation indicates that these metrics likely measure similar underlying constructs related to supplier charges and Medicare payments, suggesting redundancy and that not all may be needed for further analyses. Therefore, PCA was apply to reduce dimensionality and focus on the key components that capture the most variance, streamlining subsequent analyses.

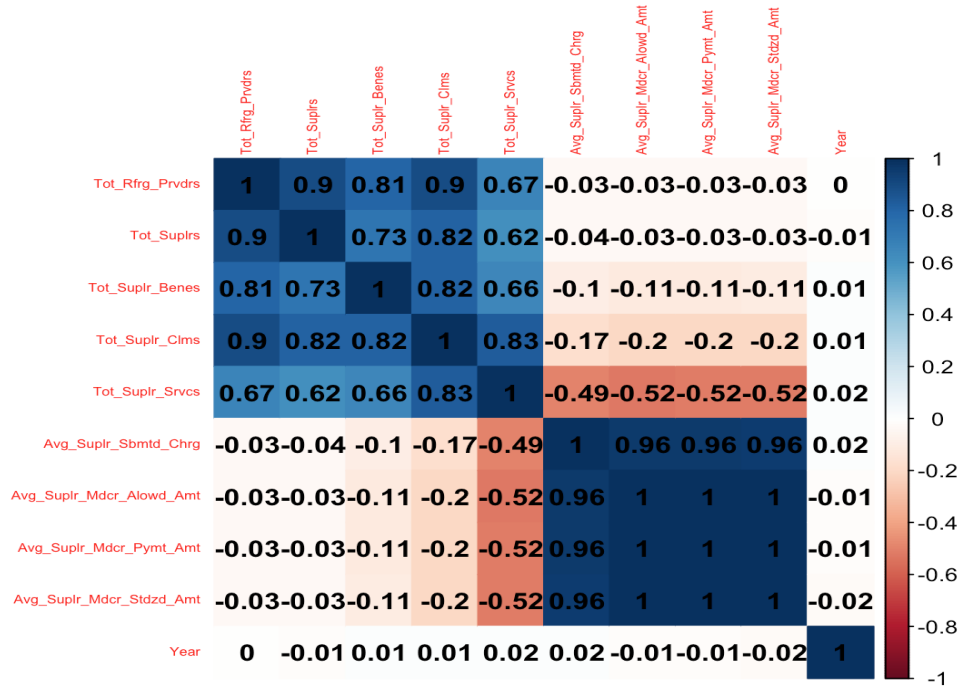

Figure 4 – Correlation Heat map

PCA Analysis

In PCA analysis, the goal was to determine the number of components necessary to retain a significant amount of variance in the dataset. The figure illustrating the cumulative explained variance versus the number of components reveals that only one component is required to explain 80% of the variance.

This finding is further supported by the loading table, the first principal component (PC_combined1) is primarily influenced by the average supplier Medicare allowed amount, payment amount, standardized amount, and submitted charge, all exhibiting strong positive loadings. The second principal component (PC_combined2) is notably shaped by the total number of referring providers and suppliers, suggesting a significant relationship between these factors and supplier behavior. The third component (PC_combined3) emphasizes the importance of total supplier services, while the fourth component (PC_combined4) also reflects a contribution from the total number of supplier claims. Additionally, the fifth and sixth components (PC_combined5 and PC_combined6) highlight a relationship with various categories of durable medical equipment and orthotic devices, alongside the year variable, indicating temporal trends in supplier activities.

This analysis confirms that dimensionality reduction can be achieved without substantial loss of information, effectively summarizing the dataset's variance. The accompanying graph visually represents these findings and enhances our understanding of the PCA results.

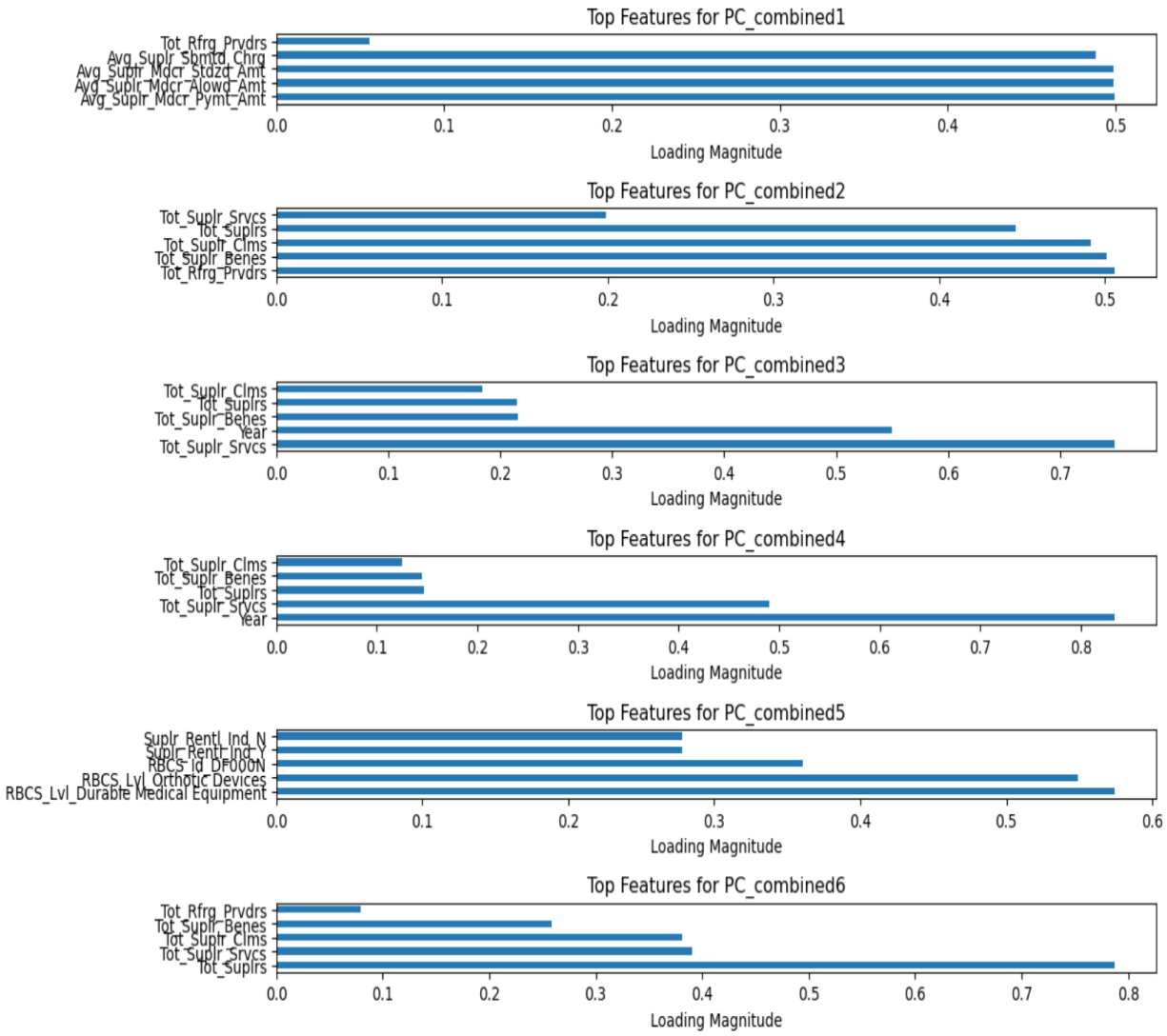

Figure 5: Cumulative Explained Variance by Principal Components

Correlation Analysis with Principal Component

To further understand the relationships in the dataset, Spearman's rank correlation coefficients were calculated and their corresponding p-values for each principal component derived from PCA (PC_combined1 to PC_combined6). The results reveal significant associations between the variance captured by these principal components and various numeric variables in the dataset,

including Avg_Suplr_Sbmtd_Chrg, Avg_Suplr_Mdcr_Pymt_Amt, Tot_Suplr_Srvcs, Tot_Rfrg_Prvdrs, and Tot_Suplr_Benes.

The first principal component (PC_combined1) exhibits strong positive correlations with variables such as Avg_Suplr_Sbmtd_Chrg, Avg_Suplr_Mdcr_Alowd_Amt, Avg_Suplr_Mdcr_Pymt_Amt, and Avg_Suplr_Mdcr_Stdzd_Amt, indicating that these variables significantly contribute to the variance captured by this component. This suggests that higher average supplier charges and Medicare payments are closely associated.

In contrast, PC_combined3 demonstrates a strong positive correlation with the variable Year (0.974480), indicating a temporal trend that may reflect changing supplier behaviors over time. Additionally, PC_combined2 shows notable positive correlations with Tot_Rfrg_Prvdrs, Tot_Suplrs, and Tot_Suplr_Benes, suggesting that the total number of referring providers and suppliers is related to this component's variance. The negative correlations found in other components, particularly for PC_combined4 and PC_combined6 with various supplier metrics, highlight underlying patterns in service provision and payment structures that warrant further investigation. Overall, this analysis underscores the potential of principal component analysis (PCA) to reveal intricate relationships among supplier-related metrics, offering valuable insights for understanding the factors influencing Medicare payments and service distribution.

The strong correlations, close to 1 or -1, signify robust relationships between the principal components and the variables, while p-values below 0.05 indicate statistically significant correlations, reinforcing the validity of these findings. This comprehensive correlation analysis

enhances our understanding of how the principal components relate to various aspects of the data, providing valuable insights for future analyses and interpretations.

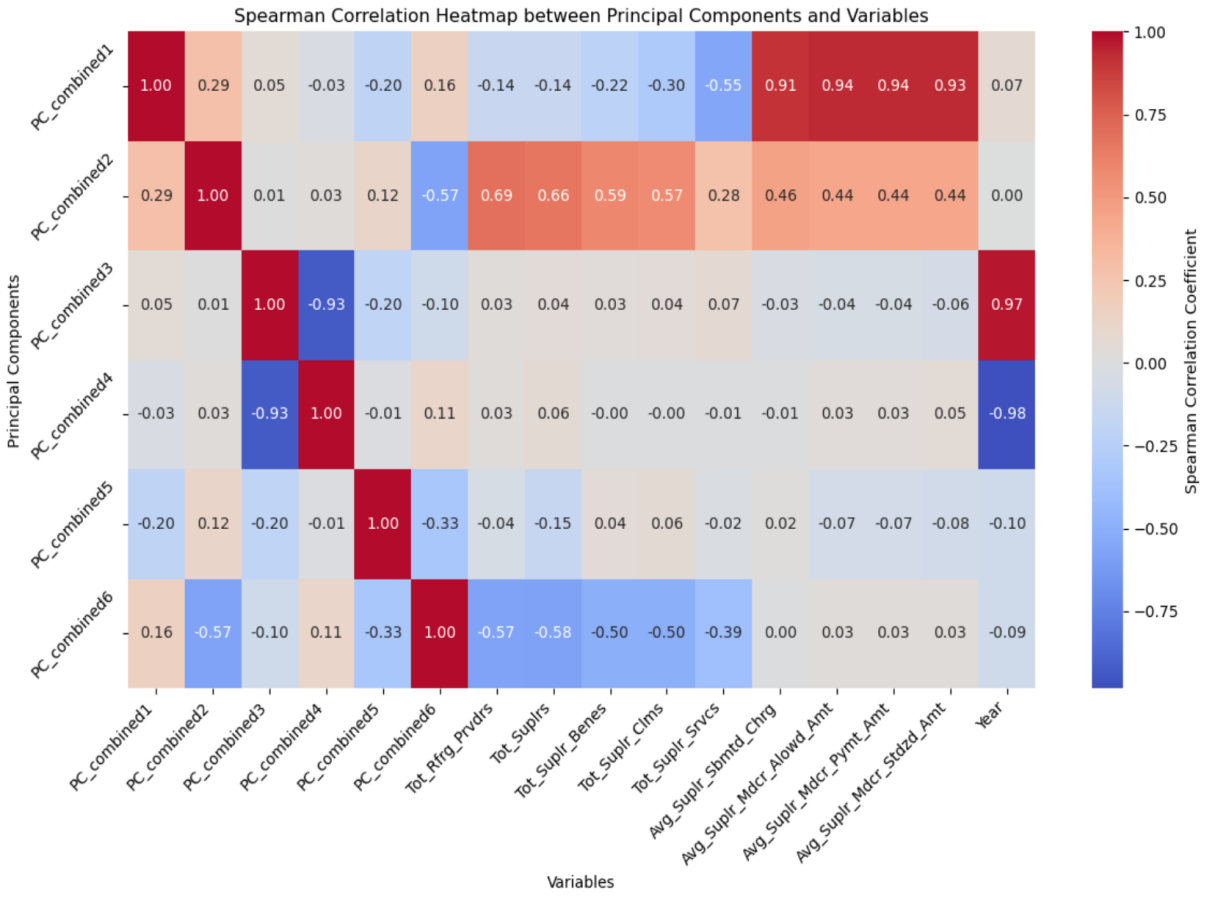

Figure 6 - Correlation Analysis with Principal Component

Outliers

In this exploratory data analysis, outliers were identified, and they will be addressed in the subsequent modeling phase using anomaly detection techniques, specifically the Isolation Forest and Local Outlier Factor (LOF) methods.

Methods

This research employs a comparative analysis approach, utilizing two models for each task to address the research questions. The focus is on detecting anomalies, classifying inefficiencies, and forecasting trends in Medicare supplier charges and service volumes over time. The models include Isolation Forest and Local Outlier Factor (LOF) for anomaly detection, Logistic Regression and Random Forest for classification, and ARIMA and SARIMA for time-series forecasting. To manage the dataset's size and complexity, Python-based tools will be used to implement and validate the models.

For anomaly detection, Isolation Forest and LOF will identify unusual patterns in supplier behavior, with k-Nearest Neighbors (kNN) applied to remove detected outliers, ensuring cleaner data for subsequent analysis. Validation will be performed using precision-recall metrics and F1-score, with the acceptance criteria set at precision, recall, and F1-score values ≥ 0.75.

For classification tasks, Logistic Regression and Random Forest will be employed, with cross-validation used to fine-tune the models and ensure robust performance. Performance will be evaluated using confusion matrices, accuracy, precision, recall, F1-score, and the Area Under the ROC Curve (AUC). The acceptance criteria for classification models are set at accuracy $\geq 80\%$, AUC ≥ 0.85, and precision, recall, and F1-score ≥ 0.75.

Time-series cross-validation analysis will be conducted using ARIMA and SARIMA to forecast trends in supplier charges and service volumes. These models will be validated through out-of-sample forecasting and residual diagnostics. Performance metrics will include Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Deviation (MAD), with the acceptance criteria being RMSE and MAE $\leq 10\%$.

To uncover associations within the dataset, the Apriori algorithm will be applied to the one-hot encoded data, which will include binary indicators for key categorical variables, such as high charges, low service volumes, underserved regions and Supplier Rental Indicator. Using the mlxtend library in Python, the analysis will involve calculating frequent item sets with a minimum support threshold of 0.05 to identify prevalent patterns. Subsequently, association rules will be generated with a confidence threshold of 0.8, enabling the exploration of relationships between different features. This approach will facilitate the identification of significant correlations, contributing valuable insights into the dynamics of Medicare supplier charges and service utilization.

Feature selection will be applied using Principal Component Analysis (PCA) and correlation techniques to reduce the number of features, ensuring that only the most relevant components are retained for the analysis. This approach simplifies the dataset by eliminating redundant or highly correlated features, which helps improve model efficiency and reduces the risk of overfitting. Following feature selection, model testing will focus on comparing the effects, stability, and efficiency of each model, using metrics such as time taken for analysis, memory usage, and overfitting. This method ensures that the models not only perform well but also operate efficiently, handling the complexity of the dataset while maintaining accuracy.

GitHub

https://github.com/Rozani1/medicare-dme-cost-analysis

References

1) Centers for Medicare & Medicaid Services. (2023). *Fiscal Year 2023 Improper Payments Fact Sheet.* https://www.cms.gov
2) Medicare.gov. What's Medicare? *U.S. Government, U.S. Centers for Medicare & Medicaid Services. The Official U.S. Government Site for Medicare.* https://www.medicare.gov/.
3) Data https://data.cms.gov/provider-summary-by-type-of-service/medicare-durable-medical-equipment-devices-supplies/medicare-durable-medical-equipment-devices-supplies-by-geography-and-service
4) Herland, M., Khoshgoftaar, T. M., & Bauder, R. A. (2018). *Big Data fraud detection using multiple Medicare data sources.* Herland et al. J Big Data (2018) https://doi.org/10.1186/s40537-018-0138-3
5) Bauder, R. A., & Khoshgoftaar, T. M. (2017). *Medicare fraud detection using machine learning methods*. 2017 16th IEEE International Conference on Machine Learning and Applications
6) Bauder, R. A., da Rosa, R., & Khoshgoftaar, T. M. (2018). I*dentifying Medicare Provider Fraud with Unsupervised Machine Learning*. 2018 IEEE International Conference on Information Reuse and Integration for Data Science
7) Shi, H., Tayebi, M. A., Pei, J., & Cao, J. (2023). *Cost-Sensitive Learning for Medical Insurance Fraud Detection With Temporal Information*. In IEEE TRANSACTIONS ON

KNOWLEDGE AND DATA ENGINEERING, VOL. 35, NO. 10, OCTOBER 2023 10451

8) Pramanik, A., Sultana, S., & Rahman, M. S. (2022). *Time Series Analysis and Forecasting of Monkeypox Disease Using ARIMA and SARIMA Model*. 2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)

9) Qu, G., Cui, S., & Tang, J. (2014). *Time Series Forecasting of Medicare Fund Expenditures Based on Historical Data—Taking Dalian as an Example*. Proceedings of the 26th Chinese Control and Decision Conference (CCDC)

Appendix A

Table 4 – Variable Descriptions (descriptions gathered from the data dictionary)

Variable	Name and Description	Data Type
Rfrg_Prvdr_Geo_Lvl	*Referring Provider Geography Level* Identifies the geographic aggregation level (State/National).	Nominal (Qualitative)
Rfrg_Prvdr_Geo_Cd	*Referring Provider Geography Code* FIPS code of the referring provider state (blank for national-level data).	Nominal (Qualitative)
Rfrg_Prvdr_Geo_Desc	*Referring Provider Geography Description* The state or region name where the provider is located (National, state, or region)	Nominal (Qualitative)
RBCS_Lvl	*Restructured BETOS Classification System Level* High-level grouping of RBCS into Durable Medical Equipment, Prosthetic and Orthotic Devices, and Drugs and Nutritional Products.	Nominal (Qualitative)
RBCS_Id	*RBCS Identifier* 6-character RBCS identifier, providing category, subcategory, and procedure information.	Nominal (Qualitative)
RBCS_Desc	*RBCS Description*	Nominal

	A concatenation of the RBCS category and subcategory description	(Qualitative)
HCPCS_Cd	*Healthcare Common Procedure Coding System Code* The HCPCS code for the DMEPOS products/services	Nominal (Qualitative)
HCPCS_Desc	*HCPCS Description* Description of the HCPCS code for the DMEPOS products/services	Nominal (Qualitative)
Suplr_Rentl_Ind	*Supplier Rental Indicator* Identifies whether the supplier claims are related to rentals (Y for Yes, N for No)	Nominal (Qualitative)
Tot_Rfrg_Prvdrs	*Total Referring Providers* The total number of referring providers ordering DMEPOS products/services	Discrete (Quantitative)
Tot_Suplrs	*Total Suppliers* The total number of suppliers rendering DMEPOS products/services	Discrete (Quantitative)
Tot_Suplr_Benes	*Total Supplier Beneficiaries* The total number of unique beneficiaries associated with the DMEPOS claims	Discrete (Quantitative)
Tot_Suplr_Clms	*Total Supplier Claims* The total number of DMEPOS claims submitted by suppliers	Discrete (Quantitative)
Tot_Suplr_Srvcs	*Total Supplier Services*	Discrete

	The total number of DMEPOS products/services rendered by suppliers	(Quantitative)
Avg_Suplr_Sbmtd_Chrg	*Average Supplier Submitted Charge* The average charge that suppliers submitted for DMEPOS products/services	Continuous (Quantitative)
Avg_Suplr_Mdcr_Alowd_Amt	*Average Supplier Medicare Allowed Amount* The average allowed amount by Medicare for the DMEPOS products/services	Continuous (Quantitative)
Avg_Suplr_Mdcr_Pymt_Amt	*Average Supplier Medicare Payment Amount* The average amount Medicare paid after deductible and coinsurance deductions	Continuous (Quantitative)
Avg_Suplr_Mdcr_Stdzd_Amt	*Average Supplier Medicare Standardized Payment Amount*	Continuous (Quantitative)

Table 5 – PICO framework

P	Population	Medicare beneficiaries receiving Durable Medical Equipment (DME) services across various geographic regions (statewide or national) in the U.S.

I	Intervention	Analysis of Medicare payments for high-supplier service volumes to identify inefficiencies.
C	Comparison	Comparison with regions where Medicare payments are within expected ranges
O	Outcome	Identification of cost inefficiencies and disparities in service delivery across different regions.
T	Time	The analysis covers the period from 2014 to 2022