Temporal Disparities in U.S. Medicare DME Costs (2014-2022): A Time-Series Cross-Validation Approach for Efficient Resource Allocation

Submitted by: Rozani Jeganathan [501282785]

Supervisor: Tamer Abdou, PhD

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 and improve resource allocation, machine learning techniques are applied to detect anomalies in supplier charges and service volumes. This study analyzes a decade of Medicare Durable Medical Equipment, Prosthetics, Orthotics, and Supplies (DMEPOS) data [3] to uncover inefficiencies, identify underserved regions, and provide actionable insights. By incorporating both supervised and unsupervised machine learning models, combined with time-series forecasting techniques, this research explores correlations between supplier behaviors, geographic disparities, and temporal trends.

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 and temporal information, offering valuable insights into geographic variations in service delivery, supplier inefficiencies, and potential inequities in Medicare's DMEPOS program.

By focusing exclusively on DMEPOS data, 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 provides a robust framework for understanding DMEPOS services and supplier behaviors across different regions, particularly underserved areas. 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. The findings demonstrate how geographic segmentation and predictive modeling can identify underserved regions, forecast resource demands, and highlight inefficiencies in high-charge, low-service areas. This comprehensive approach informs policy development by enabling Medicare to optimize costs, improve equity in healthcare access, and mitigate fraudulent activities. Ultimately, this research underscores the potential of machine learning and time-series analysis to support targeted interventions, to ensure more effective Medicare resource allocation and improve healthcare outcomes for underserved populations.

The project addresses the following research questions:

1. How do supplier service volumes and Medicare payments fluctuate over time, and which specific periods exhibit inefficiencies or potential fraud requiring targeted interventions?
2. Do certain regions consistently show higher supplier service volumes and Medicare payments, and are underserved areas disproportionately charged higher rates during specific periods, which could support targeted policy interventions?
3. How do correlations between service volumes and supplier charges evolve over time, and what temporal patterns suggest inefficiencies or irregular relationships?
4. How have supplier charges varied by region over time, and how can these temporal trends help optimize costs?

By using Python and R 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.

Despite its critical role in supporting beneficiaries, DMEPOS resource allocation remains relatively underexplored compared to other areas of Medicare spending. Regional disparities in DME spending, as highlighted by recent CMS data [10], further underscore this gap. For example, states like Texas and Oklahoma had DME spending exceeding the national average by over 10% in 2022, indicating potential inefficiencies or overutilization. Conversely, states like Alaska and Vermont spent 10% below the national average, suggesting underserved regions where beneficiaries may face barriers to accessing necessary equipment. These disparities illustrate the pressing need to address geographic inequities in Medicare's resource allocation, a core focus of this study.

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.

This study builds on insights from fraud detection research, adapting its principles to identify inefficiencies rather than solely focusing on fraudulent activities. For example, fraud detection methodologies often flag high charges and low service volumes as indicators of anomalies. Applying a similar approach, this study identifies regions where these conditions indicate broader inefficiencies in DMEPOS allocation. By incorporating regional segmentation and time-series analysis, the study bridges the gap between identifying inefficiencies and addressing geographic disparities in Medicare.

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 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 reviewed 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

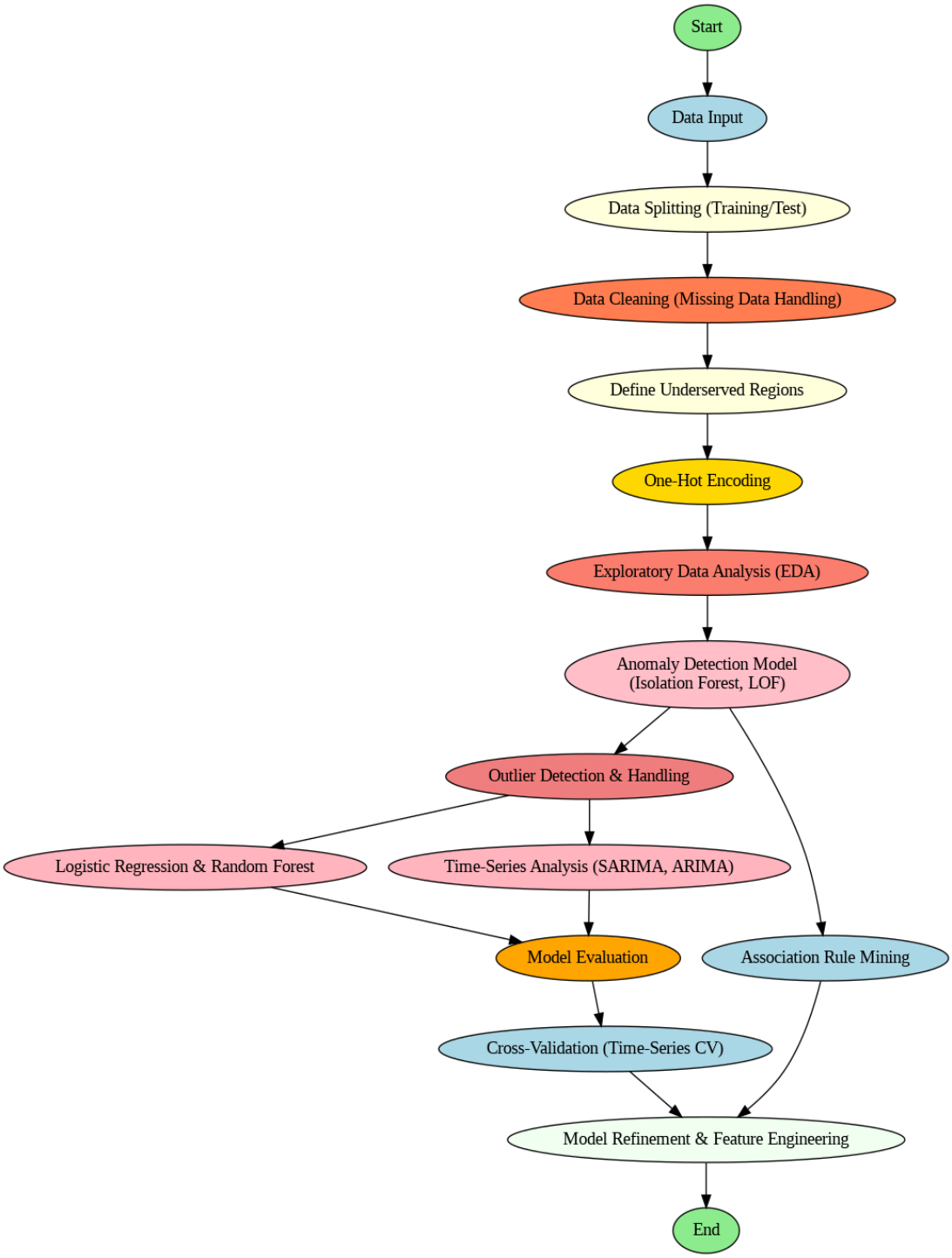


Figure 1 - Tentative overall methodology graph

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 values 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.

Refining Underserved Region Classification:

Initially, underserved regions were identified using a strict condition where regions needed to exhibit both high charges and low service volumes. However, this approach resulted in a very small subset of regions being classified as underserved, which created challenges in developing robust models due to insufficient variability in the data.

To address this limitation, the classification was revised to include regions that meet either high charges or low service volumes conditions. This broader definition resulted in:

- **High Underserved**: Regions meeting both conditions (high charges and low service volumes).

- **Low Underserved**: Regions meeting one condition (either high charges or low service volumes).
- **Not Underserved**: Regions meeting neither condition.

This refinement ensures a sufficient sample size for model training while still capturing meaningful disparities in DME service distribution.

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.

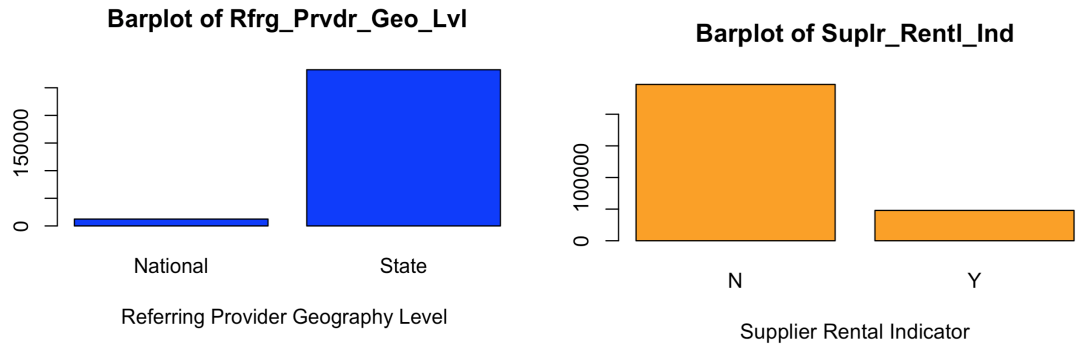


Figure 2 – Bar plot of Binary categorical variable

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 was 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 on 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.

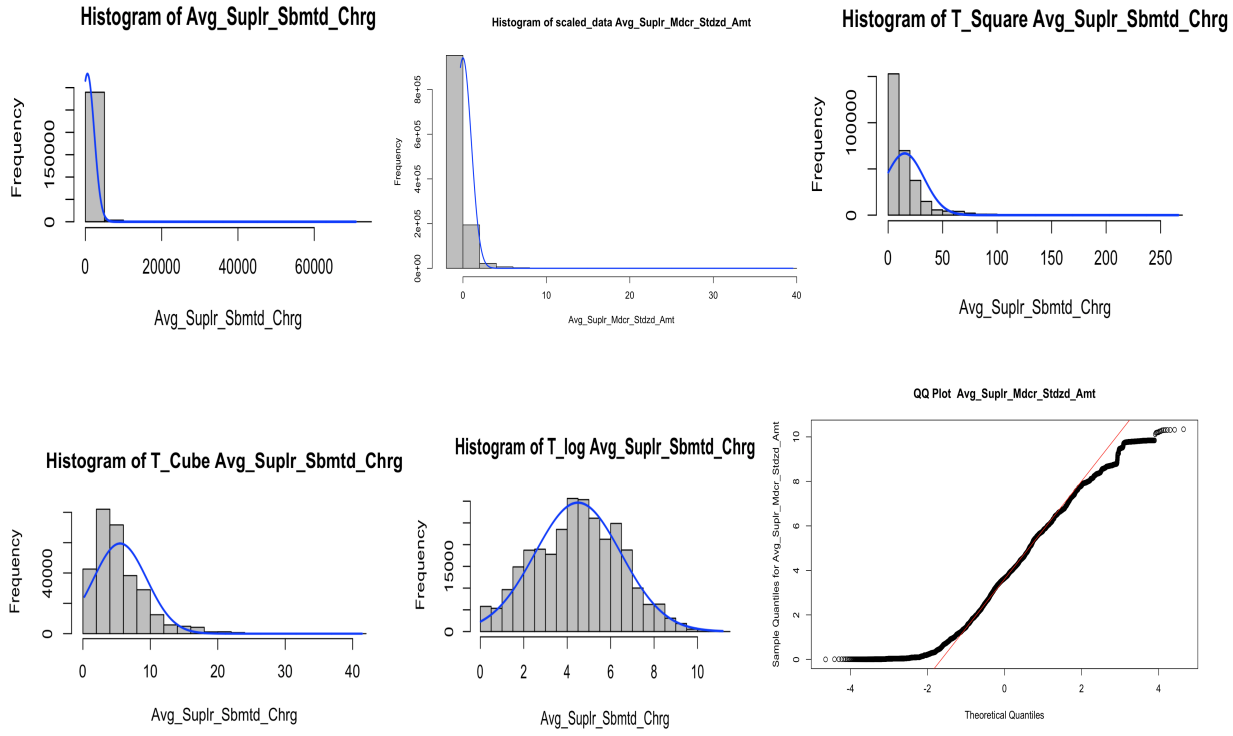


Figure 3 – Histograms and QQ Plot of Average Supplier Submitted Charge

Due to the right-skewness and non-normal distribution of most discrete variables, as confirmed by the Shapiro-Wilk test and visualized through histograms, 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. Given the non-normality of the data, Spearman's correlation coefficient was chosen to analyze the relationships between key variables. This rank-based method provides robust insights into monotonic relationships, unaffected by outliers or skewed distributions. By identifying strong correlations, such as between service volumes and charges, this analysis informs the selection of features for machine learning models aimed at detecting inefficiencies. Weak correlations, in contrast, highlight areas where nonlinear modeling or additional features may be required to capture complex dependencies.

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 continuous variables 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.

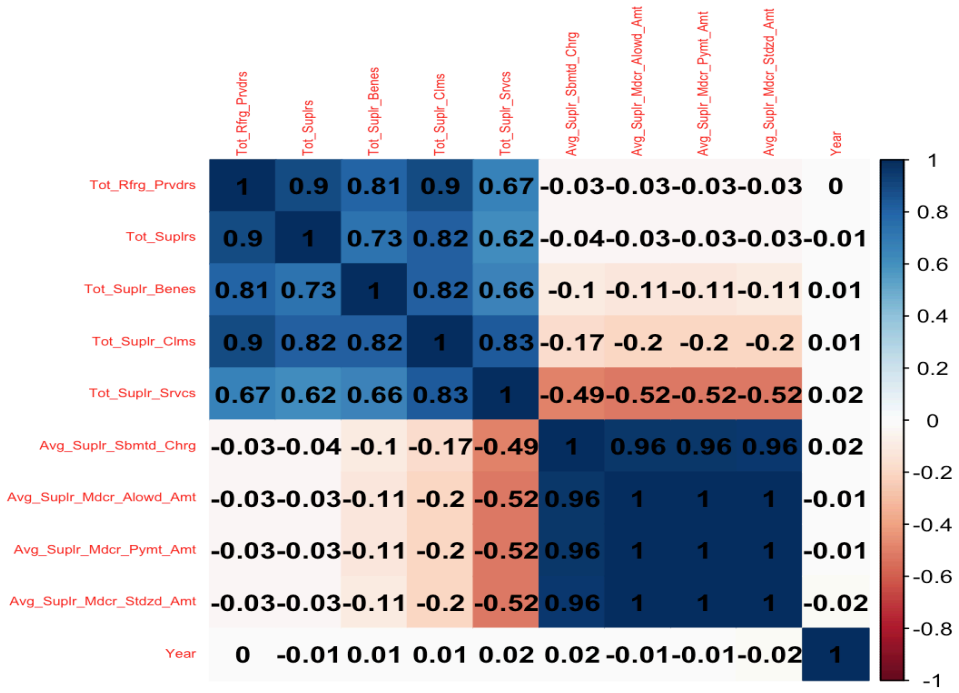


Figure 4 – Correlation Heat map

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.

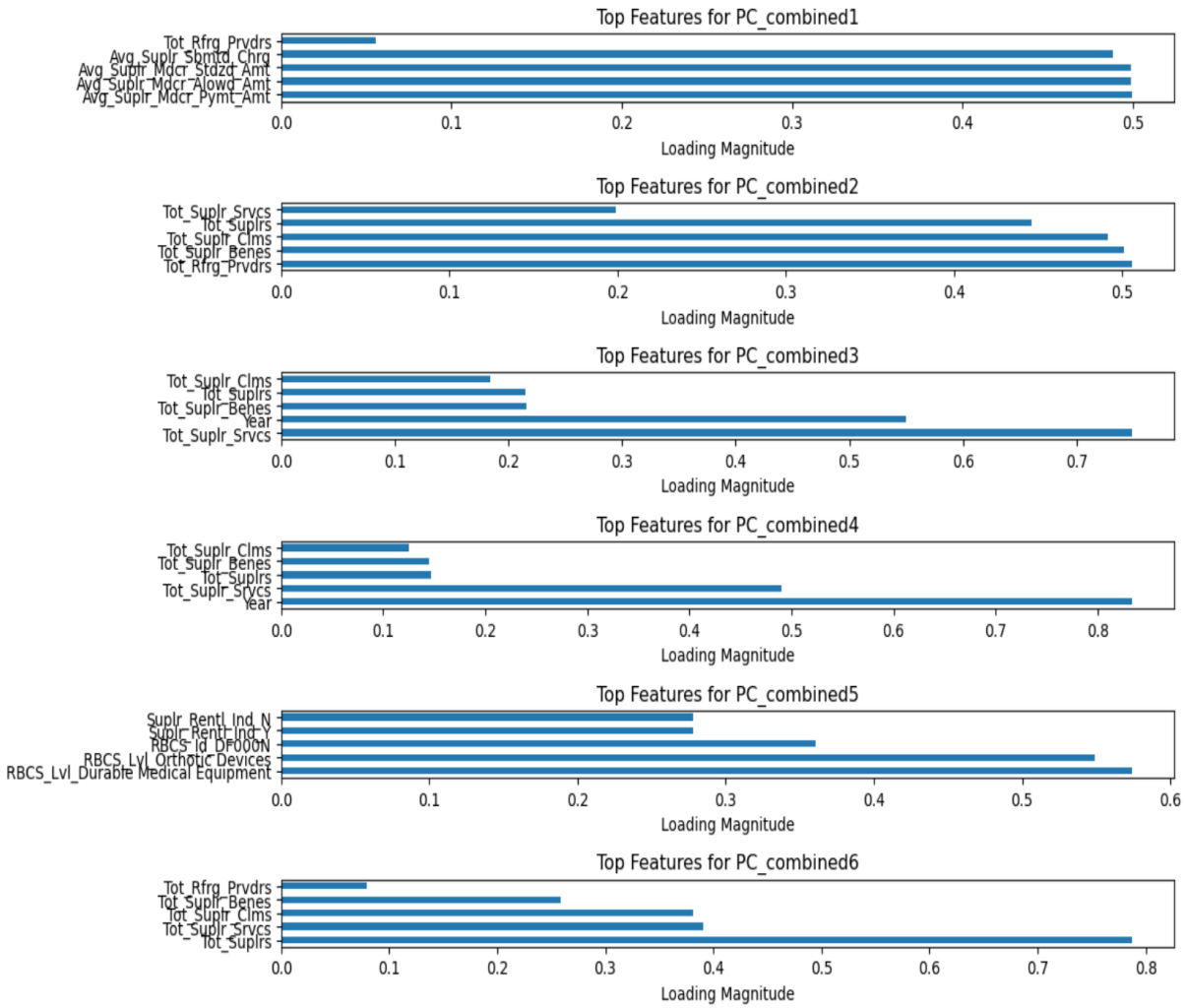


Figure 5: Cumulative Explained Variance by Principal Components

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.

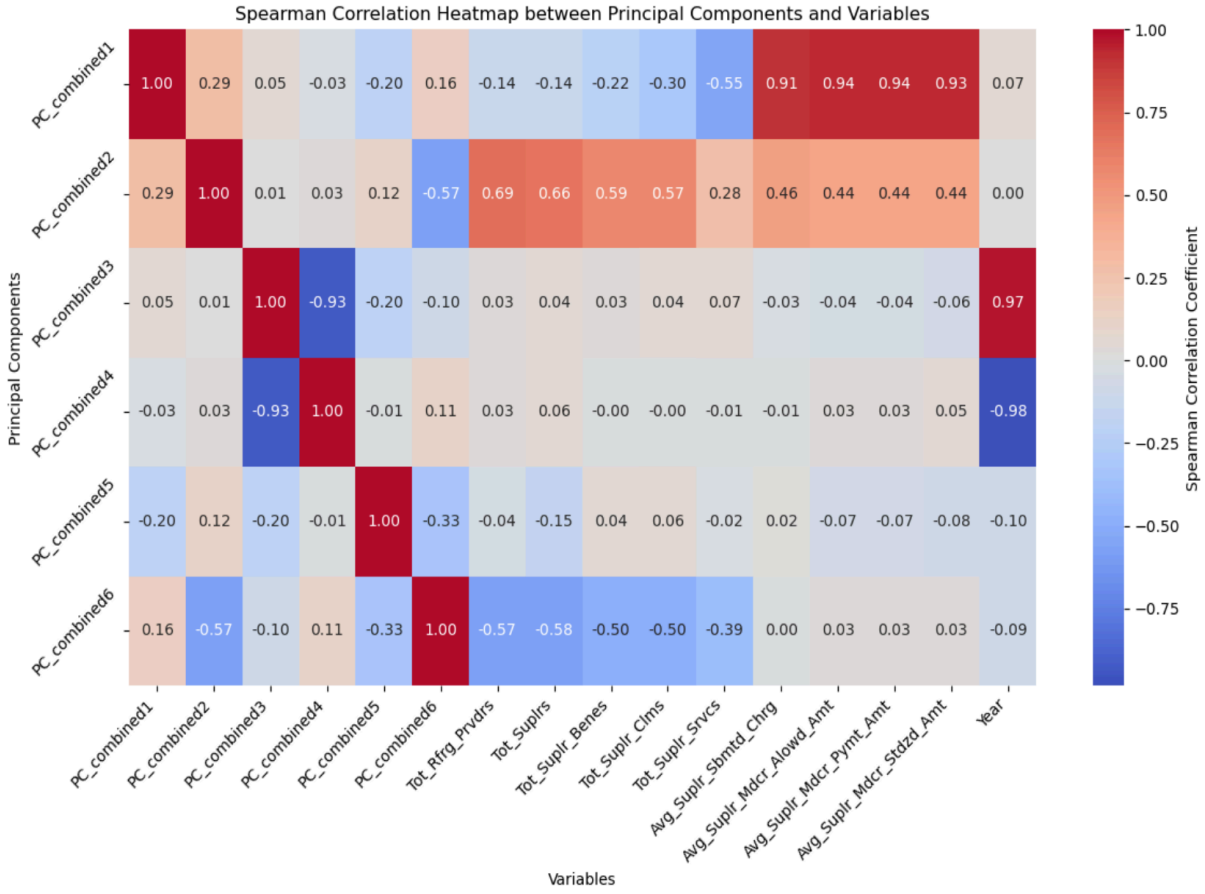


Figure 6 - Correlation Analysis with Principal Component

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 multiple 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 Prophet for time-series forecasting. Python-based tools and libraries are employed to preprocess the data, implement models, and validate results.

Anomaly Detection

For anomaly detection, Isolation Forest and LOF are used to identify unusual patterns in supplier behavior. Detected anomalies are flagged and removed to clean the dataset for subsequent analyses. Validation is performed using precision-recall metrics and F1-score, with the acceptance criteria set at precision, recall, and F1-score values ≥ 0.75. The results of anomaly detection guide the identification of regions or suppliers with outlier behavior, contributing to insights into inefficiencies and potential fraud.

Classification Tasks

For classification tasks, Logistic Regression and Random Forest are employed to categorize regions into levels of inefficiency. Cross-validation is used to fine-tune the models and ensure robust performance. Model performance is 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. These models are critical for identifying patterns in supplier and regional behavior that highlight inefficiencies and underserved areas.

Time-Series Analysis

Time-series forecasting is conducted using ARIMA and Prophet to model and predict trends in supplier charges and service volumes. These models allow for both seasonal and long-term trend analysis. Model validation includes out-of-sample forecasting and residual diagnostics, with performance evaluated using Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Squared Logarithmic Percentage (MSLP). The acceptance criteria are set at RMSE and MAE ≤ 10%. Forecasting results inform proactive resource allocation and cost optimization strategies for Medicare.

Dimensionality Reduction

Principal Component Analysis (PCA) is applied to reduce dimensionality and eliminate redundancy among highly correlated variables. The cumulative explained variance indicates that a small number of components can retain a significant portion of the dataset's variance, simplifying subsequent analyses. This step ensures that only the most relevant features are retained, improving model efficiency and reducing the risk of overfitting.

Performance Evaluation

Model performance is compared across metrics such as time taken for analysis, memory usage, and overall stability. By balancing accuracy with computational efficiency, the selected models address the complexity of the dataset while ensuring reliable and interpretable results. The findings provide actionable insights into Medicare's resource allocation and inefficiencies, particularly in underserved regions.

Results and Interpretation

This study investigates the dynamics of healthcare supplier behavior and inefficiencies using a robust mix of exploratory, predictive, and classification models. The analysis employs several advanced methodologies tailored to the specific objectives of the research:

- **Anomaly Detection Models (Isolation Forest and Local Outlier Factor)**: These unsupervised models identified outliers in service volumes and Medicare payments, highlighting periods and regions of potential inefficiency or irregular activity (Anomaly Detection and Time-Series Analysis for Supplier Service Volumes and Medicare Payments)
- **Spearman's Correlation**: Used to explore monotonic relationships between service volumes and charges, revealing key inefficiencies across regions and underserved levels (Time-Series Analysis for Correlations Between Service Volumes and Supplier Charges)
- **Supervised Classification Models (Random Forest and Logistic Regression)**: Classified regions based on their underserved levels, providing actionable insights into systemic inequities and resource allocation (Supervised Classification for Underserved Regions)
- **ARIMA and Prophet Forecasting Models**: Forecasted temporal trends in supplier charges and service volumes, enabling predictions of future inefficiencies and cost trajectories (Forecasting Average Supplier Submitted Charges Using ARIMA and Prophet) and (Forecasting Total Supplier Services Using ARIMA and Prophet)

This section systematically presents findings for each research question, supported by logical interpretations, performance evaluations of the models, and their implications for addressing inefficiencies in healthcare systems.

PCA Analysis Results

pca_important_features

Year	Principal Component	Feature	Loading
2015	PC_combined1	Avg_Suplr_Mdcr_Pymt_Amt	0.498236
2015	PC_combined1	Avg_Suplr_Mdcr_Alowd_Amt	0.498195
2015	PC_combined1	Avg_Suplr_Sbmtd_Chrg	0.492379
2015	PC_combined2	Tot_Suplrs	0.456075
2015	PC_combined2	Tot_Suplr_Benes	0.498374
2015	PC_combined2	Tot_Suplr_Srvcs	0.224193
2017	PC_combined1	Avg_Suplr_Mdcr_Pymt_Amt	0.499592
2017	PC_combined1	Avg_Suplr_Mdcr_Alowd_Amt	0.499474
2017	PC_combined1	Avg_Suplr_Sbmtd_Chrg	0.488267
2017	PC_combined2	Tot_Suplrs	0.447207
2017	PC_combined2	Tot_Suplr_Benes	0.495443
2017	PC_combined2	Tot_Suplr_Srvcs	0.229105
2019	PC_combined1	Avg_Suplr_Mdcr_Pymt_Amt	0.500441
2019	PC_combined1	Avg_Suplr_Mdcr_Alowd_Amt	0.500399
2019	PC_combined1	Avg_Suplr_Sbmtd_Chrg	0.487888
2019	PC_combined2	Tot_Suplrs	0.435662
2019	PC_combined2	Tot_Suplr_Benes	0.498188
2019	PC_combined2	Tot_Suplr_Srvcs	0.208863

Figure 7: Key PCA Features and Loadings (2015, 2017, 2019)

The PCA analysis revealed that supplier financial metrics are the primary drivers of inefficiencies in the dataset, with the first principal component (PC_combined1) explaining 80% of the variance. Metrics such as Avg_Suplr_Mdcr_Pymt_Amt, Avg_Suplr_Mdcr_Alowd_Amt,

and Avg_Suplr_Sbmtd_Chrg consistently had the highest loadings, highlighting their critical role in determining inefficiencies. The second principal component (PC_combined2) emphasized the importance of Tot_Suplrs, Tot_Suplr_Benes, and Tot_Suplr_Srvcs, indicating the relevance of service volumes and supplier counts as secondary factors. Temporal trends demonstrated consistent feature importance across years (2015, 2017, 2019), reinforcing the robustness of these findings. By reducing dimensionality while retaining 80% of the dataset's variance, PCA simplified the analysis and provided interpretable insights to inform feature selection for subsequent modeling. This approach enabled the identification of key relationships while minimizing information loss, supporting actionable recommendations to address inefficiencies in supplier charges and service delivery.

Principal Component Analysis (PCA) was selected as the dimensionality reduction technique for this study to address redundancy among highly correlated variables. By transforming the data into uncorrelated components, PCA simplifies the feature space while retaining maximum variance, ensuring that key drivers of inefficiencies are preserved. Unlike t-SNE or LDA, PCA offers interpretable components and aligns with the study's focus on identifying meaningful relationships across features to support actionable policy recommendations.

Results by Research Question

Research Question 1

How do supplier service volumes and Medicare payments fluctuate over time, and which specific periods exhibit inefficiencies or potential fraud requiring targeted interventions?

Model Evaluation

The model evaluation is done across key criteria, including Consistency, Robustness, Scalability, and Risk of Over-Detection, to assess the strengths and limitations of Isolation Forest (ISF) and Local Outlier Factor (LOF) in anomaly detection.

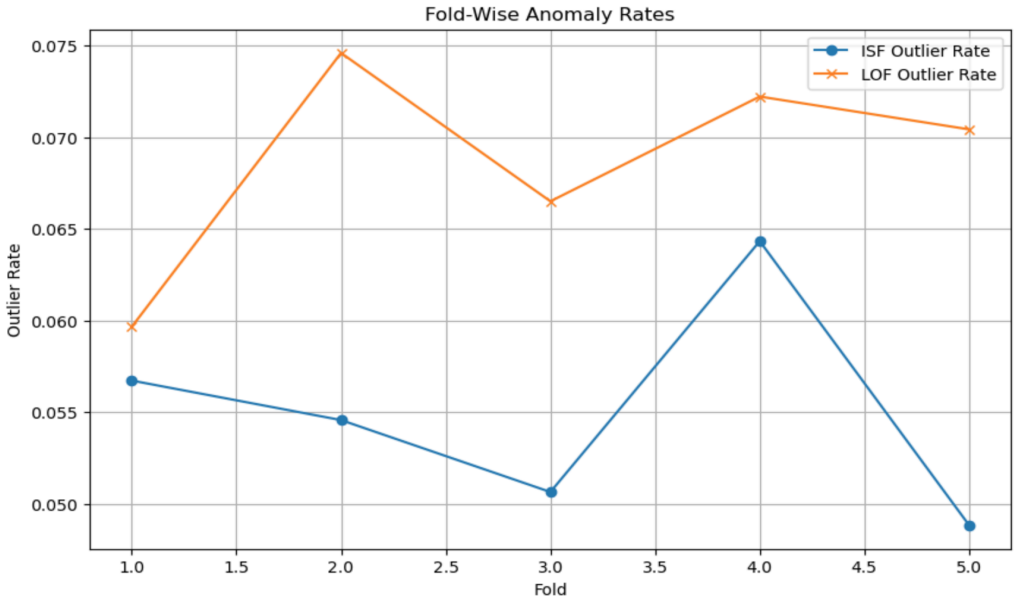


Figure 8 : Fold-wise Anomaly rates of ISF vs LOF

Fold	ISF Outlier Rate	LOF Outlier Rate	Shared Anomalies
1	0.06	0.06	876
2	0.05	0.07	799
3	0.05	0.07	677
4	0.06	0.07	893
5	0.05	0.07	659

Table 4 - Rounded ISF and LOF Comparison Table for training data

Consistency

ISF demonstrates remarkable consistency, as observed in both training and test datasets. The test outlier rate of **0.0555** aligns closely with the training average of **0.055**, indicating its ability to generalize effectively to unseen data.

Robustness

The global anomaly detection approach of ISF makes it highly robust to noisy or sparse datasets. Unlike LOF, which depends on local density variations and is sensitive to localized patterns, ISF maintains a balanced detection rate even when data distributions vary significantly. This robustness is further validated in the test results, where ISF outlier rates remain stable across folds.

Lower Risk of Over-Detection

The results underscore that LOF detects a higher rate of anomalies compared to ISF. For instance:

- Fold 2: LOF detects anomalies at a rate of 0.07 compared to ISF's 0.05.
- This tendency of LOF to over-detect can lead to an increased risk of false positives and the potential removal of valid data points. ISF's more measured detection minimizes this risk, ensuring that only the most significant anomalies are flagged.

Shared Anomalies

From a computational perspective, ISF's tree-based structure offers scalability advantages, especially in large datasets. The LOF method's reliance on local density calculations can introduce inefficiencies, particularly with larger datasets or those with uneven distribution.

Metric	ISF	LOF
Average Outlier Rate (Training)	0.055	0.066
Average Outlier Rate (Test)	0.056	0.071
Shared Anomalies (Average)	~811	~811
Risk of Over-Detection	Low	High
Robustness to Noise/Sparsity	High	Moderate
Scalability	High	Moderate

Table 5 - Comparative Performance Metrics for Isolation Forest and Local Outlier Factor

Conclusion from Model Comparisons

The results from both training and test datasets solidify ISF's position as a consistent, robust, and efficient anomaly detection model. Its balanced detection rate minimizes false positives while retaining key anomalies. In contrast, LOF’s tendency to over-detect may suit applications requiring more sensitivity but risks introducing noise and inefficiencies.

Conclusion Focused on RQ1

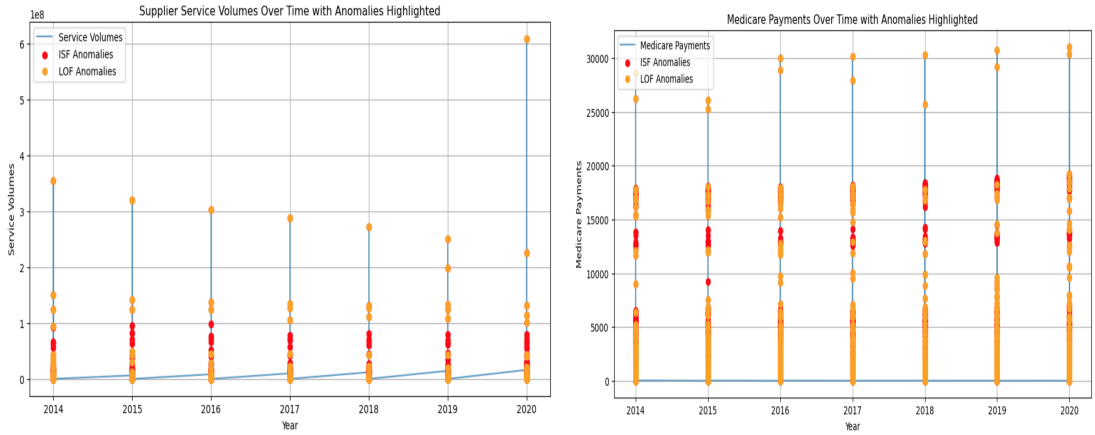


Figure 9: Feature Contributions for Anomalies vs. Normal

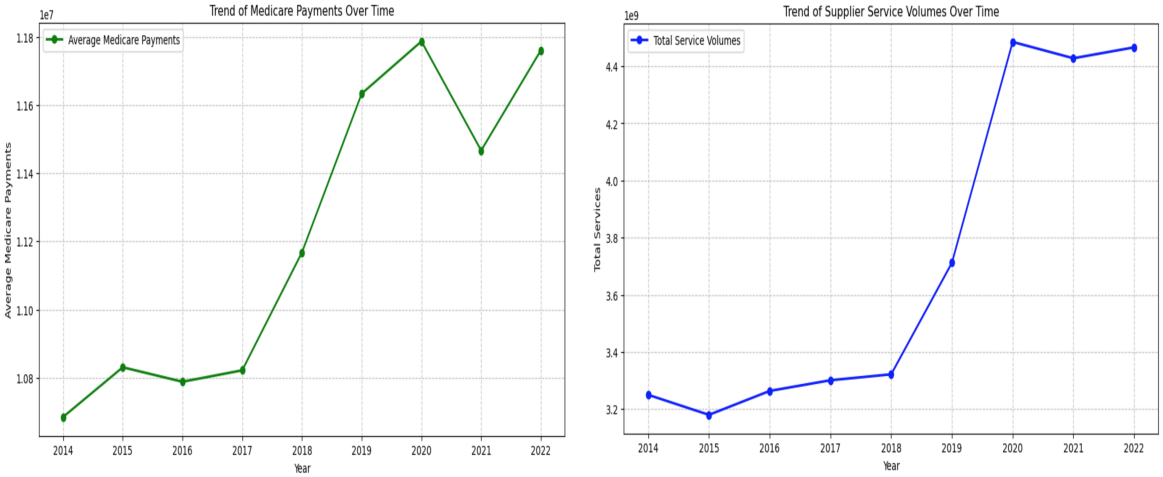


Figure 10: Trends of Medicare Payments and Supplier Service Volumes Over Time

Analyzing anomalies across service volumes and Medicare payments reveals systemic issues and localized inefficiencies. 2020 stands out as a critical year, highlighting potential external impacts and supplier irregularities. Using ISF and LOF together strengthens detection and provides actionable insights.

Temporal Trends

Service Volumes: There is a steady rise in total service volumes post-2018, reflecting increased supplier activity.

Medicare Payments: Average Medicare payments followed a relatively flat trend but showed sharp increases beginning in 2018, with a notable dip in 2021 and recovery in 2022 - High anomaly densities in both supplier service volumes and Medicare payments are observed in later indices of the dataset. These clusters indicate irregularities or extreme behaviors.

Alignment: Service volumes and Medicare payments demonstrate interconnected growth trends, with significant increases post-2018. However, anomalies identified in later periods and during 2021 (likely due to systemic disruptions such as COVID-19) indicate inefficiencies requiring further investigation.

Limitations and Overcoming Challenges in Anomaly Detection

One major limitation in the anomaly detection process was the lack of labeled anomalies, which made it impossible to use traditional metrics like precision, recall, and F1-score for validation. These metrics require ground truth labels to accurately assess how well the models identify true anomalies and avoid false positives. To address this challenge, the approach used alternative evaluation strategies tailored for an unsupervised learning context. The approach evaluated the consistency of outlier rates across cross-validation and test datasets, with the Isolation Forest (ISF) model showing high consistency, as the training outlier rate (0.055) closely matched the test rate (0.056). This consistency indicates the model's ability to generalize effectively. Additionally, the approach compared the risk of over-detection between models. The Local Outlier Factor (LOF) showed a higher tendency to over-detect, whereas ISF maintained a

balanced detection rate, minimizing false positives. Both ISF and LOF identified approximately 811 shared anomalies, enhancing the credibility of detected anomalies despite the lack of labeled data. ISF was also favored for its robustness to noise and scalability, crucial for handling large healthcare datasets. Moving forward, periodic manual reviews of flagged anomalies are recommended for model refinement, along with emphasizing explainability to help stakeholders understand model decisions. By focusing on consistency, robustness, and explainability, the approach successfully overcame the challenges posed by the lack of labeled data while extracting valuable insights from the anomaly detection models.

Policy Recommendations for RQ1

Strengthen Real-Time Monitoring and Detection

- Implement advanced anomaly detection systems, such as Isolation Forest or similar models, to enable real-time monitoring.
- Set up alerts for periods where Medicare payments deviate significantly from service volume trends.

Future Directions

- Focus audits on anomalies in later indices, particularly unusual spikes in Medicare payments unrelated to corresponding increases in service volumes.
- Investigate the causes of the 2021 dip in Medicare payments and anomalies observed in both service volumes and payments.
- Enhance anomaly detection systems to continuously monitor for irregular patterns and mitigate risks of inefficiencies or fraud.

Research Question 2

Do certain regions consistently show higher supplier service volumes and Medicare payments, and are underserved areas disproportionately charged higher rates during specific periods, which could support targeted policy interventions?

Model Evaluation

Evaluation of the machine learning models is done across three main criteria **Effectiveness**, **Efficiency**, and **Stability**.

Effectiveness

Effectiveness assesses how well each model predicts the target variable. The following performance measures are used:

AUC (Area Under the Curve) : Measures the ability of the model to distinguish between classes. Higher AUC indicates better performance in separating positive and negative class instances.

MCC (Matthews Correlation Coefficient): A balanced measure of classification quality that accounts for true positives, false positives, true negatives, and false negatives. A higher MCC indicates better model performance.

Brier Score: Measures the accuracy of probabilistic predictions. A lower Brier score indicates a better model.

Effectiveness is measured by AUC, MCC, and Brier score, reflecting how well the model differentiates between classes and makes probabilistic predictions.

Model Configuration	AUC (Training)	AUC (Test)	MCC	Brier Score	Interpretation
Random Forest (Training)	1.00	N/A	1.00	0.00	Perfect performance during training; overfitting potential.
Logistic Regression (Training)	0.90	N/A	0.64	0.05	Reasonable performance with lower MCC compared to Random Forest.
Random Forest (Original)	1.00	1.00	0.97	0.01	Highly effective with near-perfect AUC, high MCC, and low Brier score.
Logistic Regression (Original)	0.90	0.90	0.63	0.05	Effective but with lower AUC and MCC compared to Random Forest.
Random Forest (Hyperparameter-Tuned)	1.00	1.00	0.97	0.01	Effective, nearly identical to the original Random Forest model.
Logistic Regression (Hyperparameter-Tuned)	0.90	0.90	0.63	0.05	Effective, with minor MCC improvements over the original Logistic Regression.
Random Forest (Feature Selected)	1.00	1.00	0.97	0.01	Effective, similar to previous Random Forest configurations with minor changes.
Logistic Regression (Feature Selected)	0.90	0.90	0.63	0.05	Effective, with no significant performance changes from other configurations.

Table 6 - Performance Comparison of Random Forest vs Logistic Regression

Model Performance Metrics:

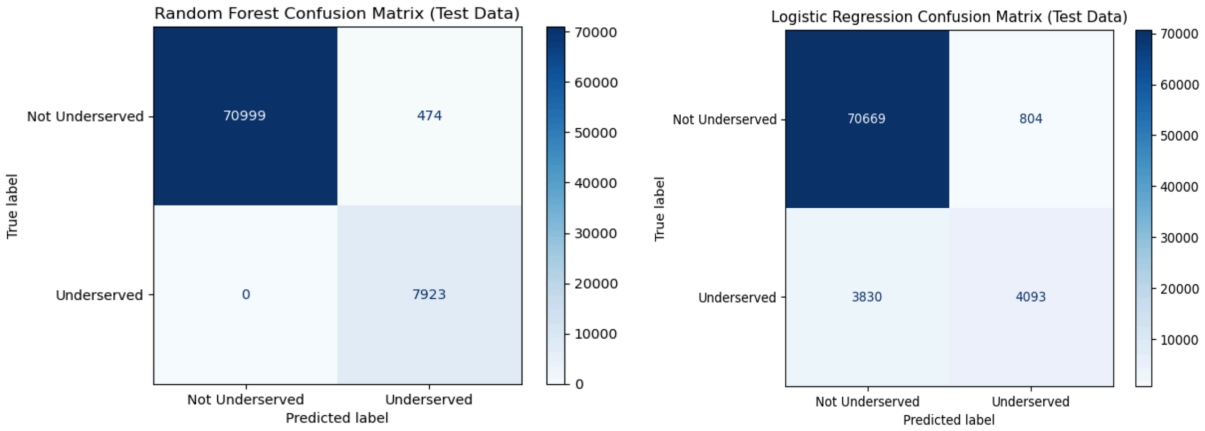


Figure 11 : Confusion Matrix Random Forest vs Logistic Regression

Model Configuration	Accuracy	Precision	Recall	F1 - Score	Interpretation
Random Forest (Training)	1.00	1.00	1.00	1.00	Perfect performance during training; likely overfitting to the training dataset.
Logistic Regression (Training)	0.94	0.94	0.94	0.93	Strong training performance, indicating a balanced ability to predict classes.
Random Forest (Original)	1.00	1.00	1.00	1.00	Near-perfect performance; robust classification with excellent recall.
Logistic Regression (Original)	0.94	0.90	0.49	0.63	Effective in precision but weak recall; struggles with identifying all positives.

Random Forest (Hyperparameter-Tuned)	1.00	1.00	1.00	1.00	Hyperparameter tuning maintained perfect performance, with no sign of overfitting.
Logistic Regression (Hyperparameter-Tuned)	0.94	0.90	0.52	0.63	Marginal improvement in recall; overall performance remains steady.
Random Forest (Feature Selected)	1.00	1.00	1.00	1.00	Feature selection preserved Random Forest's high effectiveness and robustness.
Logistic Regression (Feature Selected)	0.94	0.90	0.52	0.63	No significant performance change; struggles to improve recall or F1-score.

Table 7 - Performance Metrics of Random Forest vs Logistic Regression

Efficiency

Efficiency measures how computationally resource-effective the models are, focusing on the time and memory consumption during training and testing:

1. **Fit Time:** Time taken to train the model.
2. **Score Time:** Time taken to test or predict with the trained model.
3. **Memory Usage:** The memory consumed by the model during training.

These metrics help assess whether the model can deliver good performance without requiring excessive computational resources.

Model Configuration	Fit Time (Avg)	Score Time (Avg)	Memory Usage (MB)	Interpretation
Random Forest (original)	13.0 s	0.26 s	376-407 MB	**Moderately efficient**, higher memory usage but reasonable training time.
Random Forest (Hyperparameter-Tuned)	12.0 s	0.25 s	407 MB	**Similar efficiency** as the base Random Forest model, with slightly lower fit time.
Random Forest (Feature Selected)	12.5 s	0.25 s	353 MB	**Slightly more efficient** with reduced memory usage compared to the other configurations.
Logistic Regression (original)	1.5 s	0.04 s	357-388 MB	**More efficient**, faster training and lower memory usage compared to Random Forest.
Logistic Regression (Hyperparameter Tuned)	1.3 s	0.04 s	357 MB	**Slightly more efficient**, with a small reduction in training time.
Logistic Regression (Feature Selected)	1.2 s	0.04 s	357 MB	**Most efficient**, like the hyperparameter-tuned version with minimal increase in memory usage.

Table 8 - Model Efficiency Comparison (Fit Time, Score Time, Memory Usage)

Stability

Stability ensures that the model performs consistently across different data splits and configurations:

1. **Cross-Validation Consistency:** Stability was assessed by performing cross-validation (5-fold) and observing how stable the performance metrics were across different splits.

2. **Overfitting or Underfitting:** Stability was also evaluated by comparing performance on training and test datasets to detect any overfitting or underfitting issues.

Stability is determined by the consistency of performance across different cross-validation splits and configurations (hyperparameter tuning, feature selection).

Model Configuration	Stability (CV Consistency)	Interpretation
Random Forest (original)	**Very stable**, consistent high AUC and MCC across splits.	**Highly stable**, consistent performance in AUC, MCC, and Brier score.
Random Forest (Hyperparameter-Tuned)	**Very stable**, similar to the base model in AUC and MCC.	**Highly stable**, almost identical to the original Random Forest model, with consistent results across different splits.
Random Forest (Feature Selected)	**Very stable**, with slight fluctuations in Brier	**Stable**, with only minor changes in performance (slight increase in Brier score).
Logistic Regression (original)	**Stable**, but fluctuations in recall across splits.	**Stable**, though the recall metric varies slightly, leading to minor inconsistencies in the model's performance across splits.
Logistic Regression (Hyperparameter Tuned)	**Stable**, with improved recall compared to the base model.	**Stable**, with better performance (improved MCC), but slight variations in recall and precision.
Logistic Regression (Feature Selected)	**Stable**, with small fluctuations.	**Stable**, similar to the hyperparameter-tuned version, with very minor changes in recall but still robust across splits.

Table 9 - Model Stability Comparison

Summary Comparison Table

Criterion	Random Forest (Test)	Random Forest (Hyperparameter-Tuned)	Random Forest (Feature-Selected)	Logistic Regression (Test)	Logistic Regression (Hyperparameter-Tuned)	Logistic Regression (Feature-Selected)
Effectiveness	Highly effective	Highly effective	Effective	Effective	Effective	Effective
AUC (Test)	1.00	1.00	1.00	0.90	0.90	0.90
MCC (Test)	0.97	0.97	0.97	0.63	0.63	0.63
Brier Score (Test)	0.01	0.01	0.01	0.05	0.05	0.05
Efficiency	Moderately efficient	Similar efficiency	Slightly more efficient	More efficient	Slightly more efficient	Most efficient
Fit Time (Avg.)	13.0 s	12.0 s	12.5 s	1.5 s	1.3 s	1.2 s
Score Time (Avg.)	0.26 s	0.25 s	0.25 s	0.04 s	0.04 s	0.04 s
Memory Usage (MB)	376-407	407	353	357-388	357	357
Stability	Highly stable	Highly stable	Stable	Stable	Stable	Stable

Table 10 - Summary Comparison Table

Conclusion from Model Comparisons

Random Forest consistently met or exceeded all acceptance criteria across all configurations, indicating excellent performance in classifying regions into levels of inefficiency. The near-perfect accuracy, AUC, precision, recall, and F1-score make Random Forest a suitable model for this classification task.

Logistic Regression, while showing good performance in accuracy and precision, failed to meet the acceptance criteria for recall and F1-score. This indicates that Logistic Regression struggles to identify all positive instances, which could lead to underserved regions being misclassified or overlooked. This limitation suggests that Logistic Regression may not be as reliable as Random Forest for this specific task, particularly where recall is critical for identifying inefficiencies.

Random Forest performs excellently in terms of effectiveness, but it is less efficient than Logistic Regression. The hyperparameter tuning and feature selection bring only small changes in effectiveness and efficiency for Random Forest. Logistic Regression is more efficient, especially in the hyperparameter-tuned and feature-selected versions, but its effectiveness (AUC and MCC) is consistently lower than Random Forest. Feature selection slightly reduces memory usage and fit time across both models but does not significantly impact effectiveness. Hyperparameter tuning leads to small improvements in recall and MCC for Logistic Regression without significant changes to overall performance.

Mann-Whitney U Test Results

To validate the differences in Medicare payments and supplier service volumes between underserved and not underserved regions, Mann-Whitney U tests were conducted. These tests are suitable for non-normally distributed data and provide robust comparisons of medians.

Medicare Payments

- **U-statistic**: 87,769,157.00
- **P-value**: 0.0000
- **Interpretation**:
    - The significant p-value indicates a statistically significant difference in the distributions of Medicare payments.
    - Underserved regions exhibit higher and more variable Medicare payments compared to not underserved regions.

Supplier Services

- **U-statistic**: 498,318,268.00
- **P-value**: 0.0000
- **Interpretation**:
    - The significant p-value confirms a substantial difference in supplier services.
    - Underserved regions have significantly fewer supplier services, highlighting critical disparities in healthcare access.

Conclusion Focused on RQ2

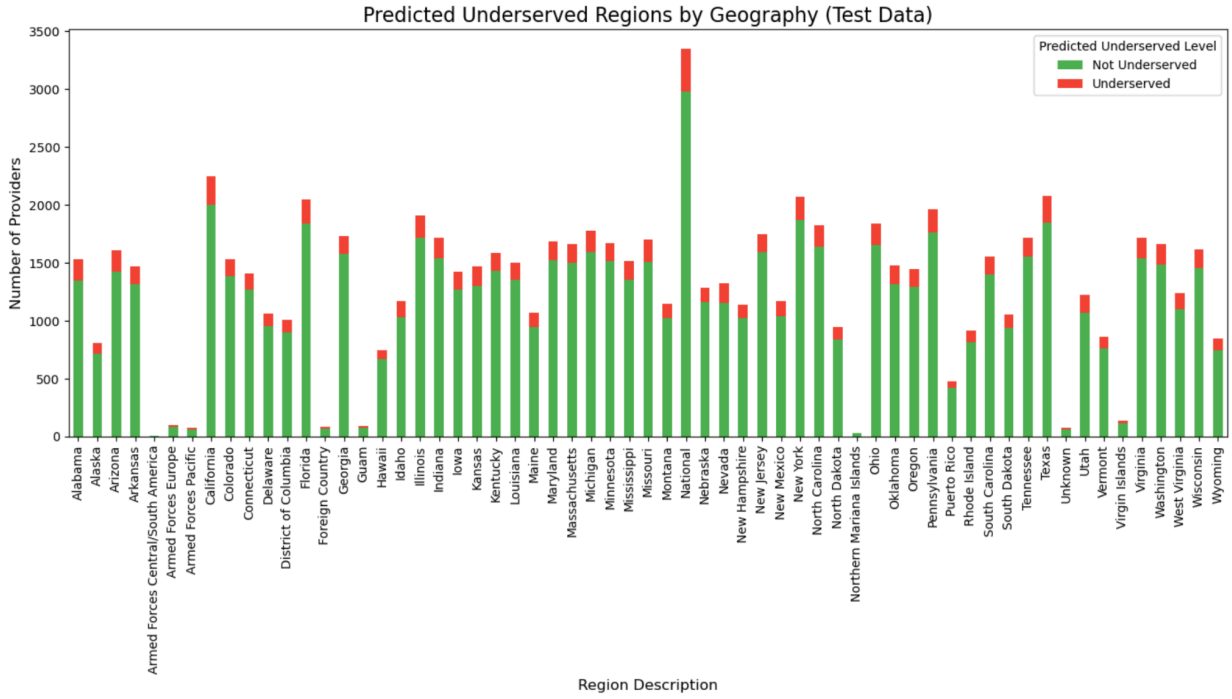


Figure 12 : Predicted Underserved regions by Geography

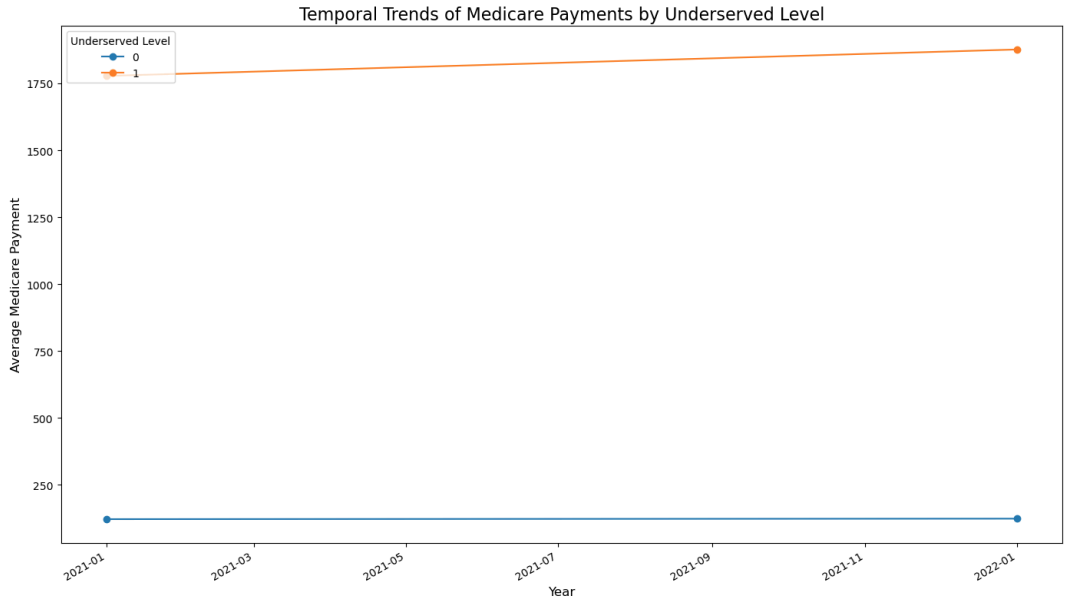


Figure 13 : Temporal trends of Medicare payments by level

The analysis addressing **RQ2** aimed to explore the relationships between service volumes and supplier charges across regions and underserved levels, providing critical insights into disparities and inefficiencies. The following conclusions are derived from the conducted analysis:

1. **Temporal Patterns**:
    a. Temporal trends showed that in well-served regions, relationships remained consistent over time.
    b. In contrast, underserved regions exhibited declining relationships between service volumes and charges, signaling worsening inefficiencies over the analyzed period.

2. **Regional Disparities:**
    a. Well-served regions like California and Texas consistently showed higher service volumes and Medicare payments.
    b. Underserved regions, including rural areas, exhibited lower service volumes but disproportionately higher Medicare payment rates, particularly in 2018 and 2019.

3. **Dimensionality Reduction (PCA)**:
    a. PCA highlighted supplier financial metrics (Avg_Suplr_Sbmtd_Chrg, Avg_Suplr_Mdcr_Pymt_Amt) as dominant contributors to inefficiencies, aligning with findings from correlation analysis.
    b. Temporal consistency in feature importance across years reinforced the robustness of these relationships.

4. **Insights on Inefficiencies**:
    a. The weaker correlations and declining temporal trends in underserved regions underscore inefficiencies likely driven by resource constraints or inequitable supplier practices.
    b. Anomalies in specific periods (e.g., spikes in 2017 and 2019 flagged by anomaly detection models) further support the presence of irregularities requiring targeted interventions.

Policy Recommendations for RQ2

- Addressing inefficiencies in underserved regions requires targeted cost optimizations, improved resource allocation, and better monitoring of supplier practices.
- Policies should focus on stabilizing the relationship between service volumes and charges to ensure equitable healthcare delivery.

Future Directions

- Incorporating additional socio-economic and demographic variables could enhance the interpretability and effectiveness of models in explaining disparities.
- Further exploration of non-linear models may uncover complex dynamics not captured by Spearman's correlation or Logistic Regression.

Research Question 3

How do correlations between service volumes and supplier charges evolve over time, and what temporal patterns suggest inefficiencies or irregular relationships?

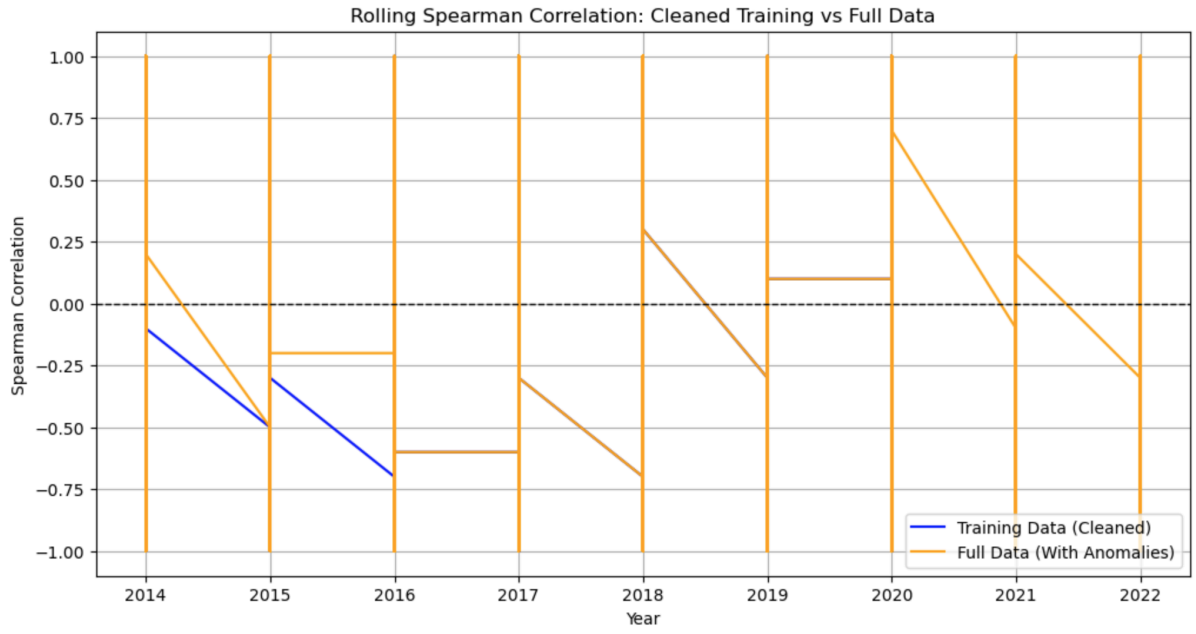


Figure 14 : Rolling Spearman Correlation Between Cleaned Training Data and Full Data

The rolling correlation was used instead of a standard correlation to capture the temporal dynamics and changes in the relationship between service volumes (Tot_Suplr_Srvcs) and submitted charges (Avg_Suplr_Sbmtd_Chrg) over time. The analysis applied a 5-year rolling window to assess how closely these two variables are related, indicating potential shifts in the efficiency or behavior of suppliers. The analysis was performed on both the cleaned dataset (anomalies removed) and the full dataset (including anomalies). The comparison helped identify periods of stability or instability in correlations and revealed how anomalies may have influenced the overall relationship between charges and services. These rolling correlations

allowed the detection of temporal patterns that could highlight inefficiencies, providing a more dynamic perspective on supplier behavior over time.

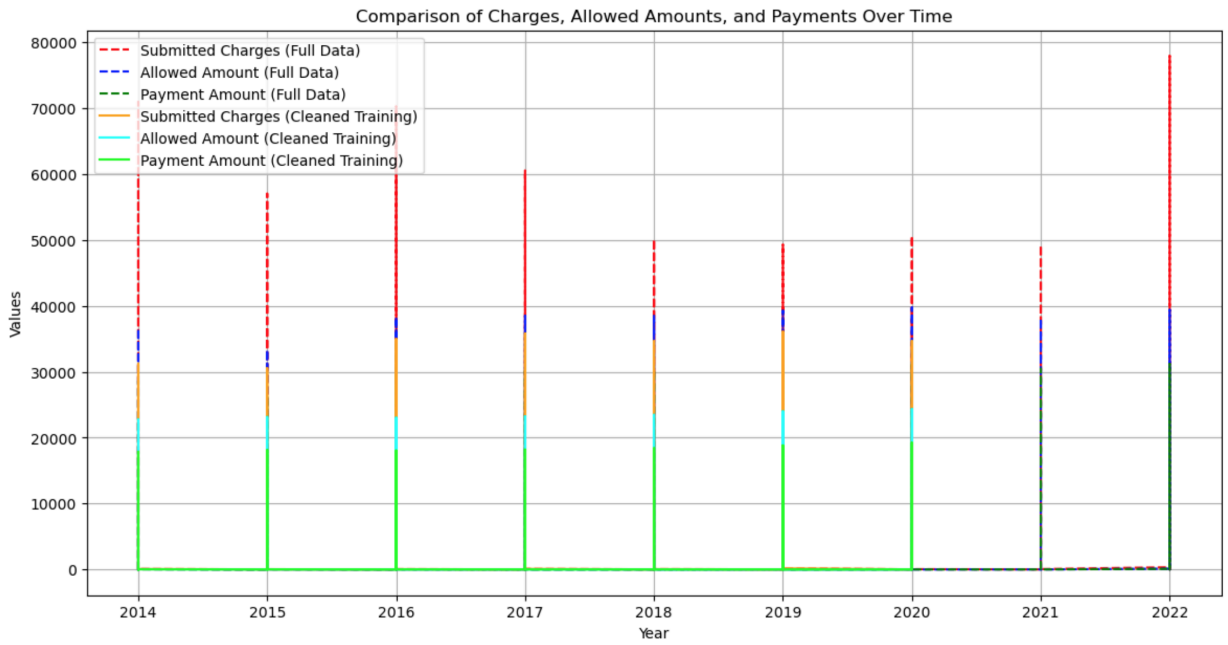


Figure 15 : Comparison of Charges, Allowed Amounts, and Payments Over Time

The comparison of submitted charges, allowed amounts, and payments over time reveals key trends in healthcare billing practices. Submitted charges are consistently higher than allowed and payment amounts, with a larger discrepancy observed in the full dataset, likely due to anomalies or irregular supplier behavior. The cleaned data shows more stable trends, providing a reliable baseline for understanding typical practices. However, significant deviations are evident in the full dataset for 2021-2022, indicating potential inefficiencies or disruptions during this period, possibly linked to external factors affecting the healthcare system.

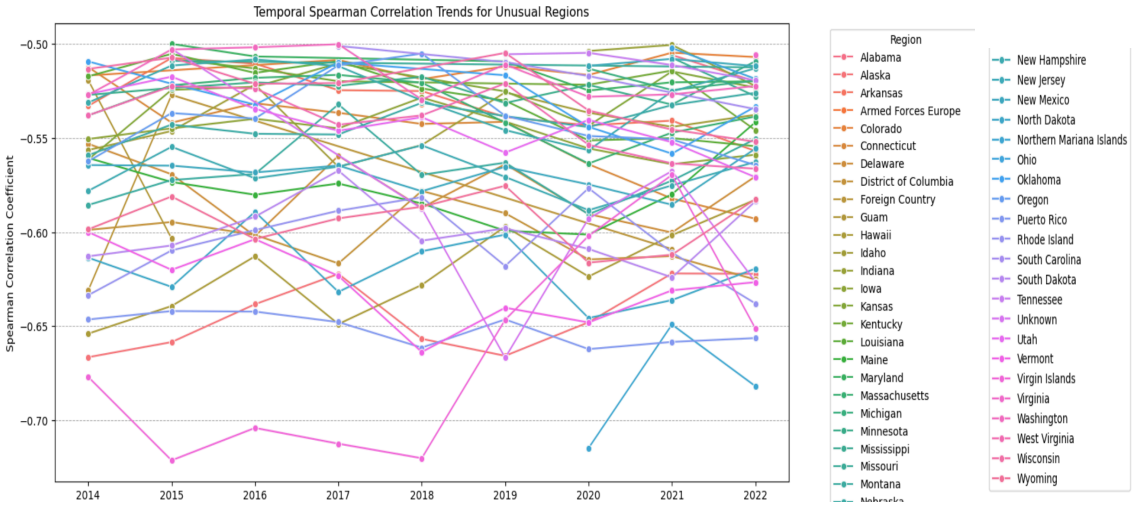


Figure 16 : Temporal Spearman correlation trends

The temporal Spearman correlation trends for unusual regions highlight important insights into the relationship between service volumes and submitted charges. Most regions exhibit stable, predominantly negative correlations ranging from -0.3 to -0.7, indicating persistent systemic inefficiencies where charges are not proportional to service volumes. While many regions show temporal stability, others experience significant fluctuations, especially post-2020, likely reflecting localized disruptions such as the impact of COVID-19. These findings point to regional disparities and systemic irregularities that warrant further investigation to address inefficiencies and improve the proportionality of charges to service volumes.

Conclusion Focused on RQ3

- **Discrepancy Between Charges and Payments**: There is a persistent discrepancy between submitted charges, Medicare allowed amounts, and actual payments. Submitted charges are consistently higher than what Medicare allows or pays, which is typical of healthcare billing practices but also indicates areas where billing efficiency might be improved.

- **Negative Correlation Trends**: Many regions exhibit predominantly negative correlations between service volumes and charges, particularly between 2020 and 2022. This implies inefficiencies in the healthcare system, as higher volumes do not correspond to proportionate charges, suggesting possible systemic issues such as regulatory impacts or supply chain disruptions.
- **Temporal Instabilities and Anomalies**: Regions with significant fluctuations, especially during 2020-2022, highlight potential disruptions due to external factors like COVID-19. These fluctuations suggest the need for targeted investigation and policy intervention to address inefficiencies and stabilize healthcare costs.

Policy Recommendations for RQ3

Based on the findings from RQ3 regarding the correlations between service volumes and submitted charges, the following policy recommendations are proposed:

- **Enhance Billing Oversight**: Since the negative correlations suggest inefficiencies where charges do not match service volumes, stricter oversight on billing practices is recommended. This could include clearer guidelines on how charges should align with service volumes.
- **Targeted Interventions in Fluctuating Regions**: Regions with significant fluctuations, especially during 2020-2022, need targeted interventions. Policymakers should investigate the reasons for these disruptions, such as the impact of COVID-19, and develop localized strategies to address these inefficiencies.

- **Incentivize Efficiency in High-Volume Services**: The negative correlations imply that high service volumes are not leading to lower costs per service. Introducing incentives for healthcare providers to improve cost-efficiency at high volumes could help align submitted charges with expected efficiencies.
- **Data-Driven Monitoring**: Establishing continuous, data-driven monitoring of service volumes and charges across regions can help identify inefficiencies early. This would allow policymakers to respond quickly, adjusting regulations or providing targeted support when discrepancies arise.

Research Question 4

How have supplier charges varied by region over time, and how can these temporal trends help optimize costs?

Model Evaluation

Metric	ARIMA (0,1,1)	Prophet	Interpretation
RMSE (Train)	1718.82	1737.69	ARIMA performs better on training data with a slightly lower RMSE.
RMSE (Test)	2211.22	1737.69	Prophet generalizes better on test data with a significantly lower RMSE. Note that Prophet's RMSE is the same for both training and test, suggesting consistent model performance across these phases.

MAE (Train)	534.26	689.65	ARIMA shows lower mean error during training, indicating better fit to training data.
MAE (Test)	614.74	689.65	ARIMA provides more accurate predictions on unseen data based on MAE. Prophet's MAE remains consistent for both training and test data, indicating uniform model behavior.
MSLP (Train)	2.54	3.32	ARIMA has a lower MSLP during training, indicating better handling of proportional errors.
MSLP (Test)	2.15	2.72	ARIMA demonstrates better generalization on test data with a lower MSLP.

Table 11 - Arima vs Prophet Comparison of Average supplier submitted charges

Metric	ARIMA (1,1,3)	Prophet	Interpretation
RMSE (Train)	1,116,137.38	909066.10	ARIMA performs better on training data with a slightly lower RMSE.
RMSE (Test)	4,287,196.02	909,066.10	Prophet generalizes better on test data with a significantly lower RMSE. Note that Prophet's RMSE is the same for both training and test, suggesting consistent model performance across these phases.
MAE (Train)	417,908.01	106909.41	ARIMA shows lower mean error during training, indicating better fit to training data.
MAE (Test)	115,423.03	106,909.41	ARIMA provides more accurate predictions on unseen data based on MAE. Prophet's MAE remains consistent for both training and test data, indicating uniform model behavior.

MSLP (Train)	4.09	8.38	ARIMA has a lower MSLP during training, indicating better handling of proportional errors.
MSLP (Test)	3.81	6.74	ARIMA demonstrates better generalization on test data with a lower MSLP.

Table 12 - Arima vs Prophet Comparison of Total supplier services

Conclusion from Model Comparisons

The results indicate that neither ARIMA nor Prophet models meet the specified acceptance criteria for RMSE and MAE being ≤ 10%. Specifically, ARIMA tends to perform better on training data with lower errors, indicating potential overfitting. However, its generalization to test data is weaker, leading to higher error metrics. Prophet, while demonstrating consistent performance across training and test data, still shows RMSE and MAE values that exceed acceptable limits, particularly due to the presence of anomalies and non-linear trends in the data.

The comparison between the ARIMA and Prophet models for both charges and services reveals key insights into their respective strengths and weaknesses:

Generalization and Consistency: Prophet generally demonstrates better generalization, as indicated by its consistent performance across both training and test datasets. For both charges and services, Prophet's RMSE and MAE remain similar, highlighting its robustness when dealing with unseen data. In contrast, ARIMA shows higher RMSE and MAE values for test data compared to training data, suggesting a tendency to overfit the training data, particularly in the service volume analysis.

Model Performance on Training Data: ARIMA tends to perform better on training data in some metrics, such as MSLP, indicating effective handling of proportional errors during training. However, this benefit does not translate well to unseen data, resulting in weaker generalization performance. Prophet's slightly higher MSLP during training suggests it has more difficulty capturing the smaller variations in training data, yet this does not negatively impact its performance on new data.

Ability to Handle Anomalies: The test datasets contain anomalies, and Prophet's ability to maintain lower error metrics under these conditions shows it is more resilient in handling irregular data patterns compared to ARIMA. ARIMA, although effective during training, struggles more with the anomalies present in the test datasets, resulting in higher errors and reduced reliability.

Practical Implications: Prophet is a more reliable choice for generalizing across different datasets, especially when consistency between training and testing performance is required. Its strength lies in capturing non-linear trends and maintaining stable performance in the presence of anomalies. ARIMA may be more suitable for scenarios where short-term accuracy during training is prioritized, but caution should be taken regarding its generalization capacity, particularly in the presence of anomalies.

Summary: Prophet is better suited for tasks requiring consistent, robust performance across both training and unseen data, while ARIMA might be favored for achieving tighter fits to historical data at the expense of generalization. The choice between these models should be based on the specific context, with an emphasis on the importance of stability and handling of anomalies for future predictions.

Conclusion Focused on RQ4

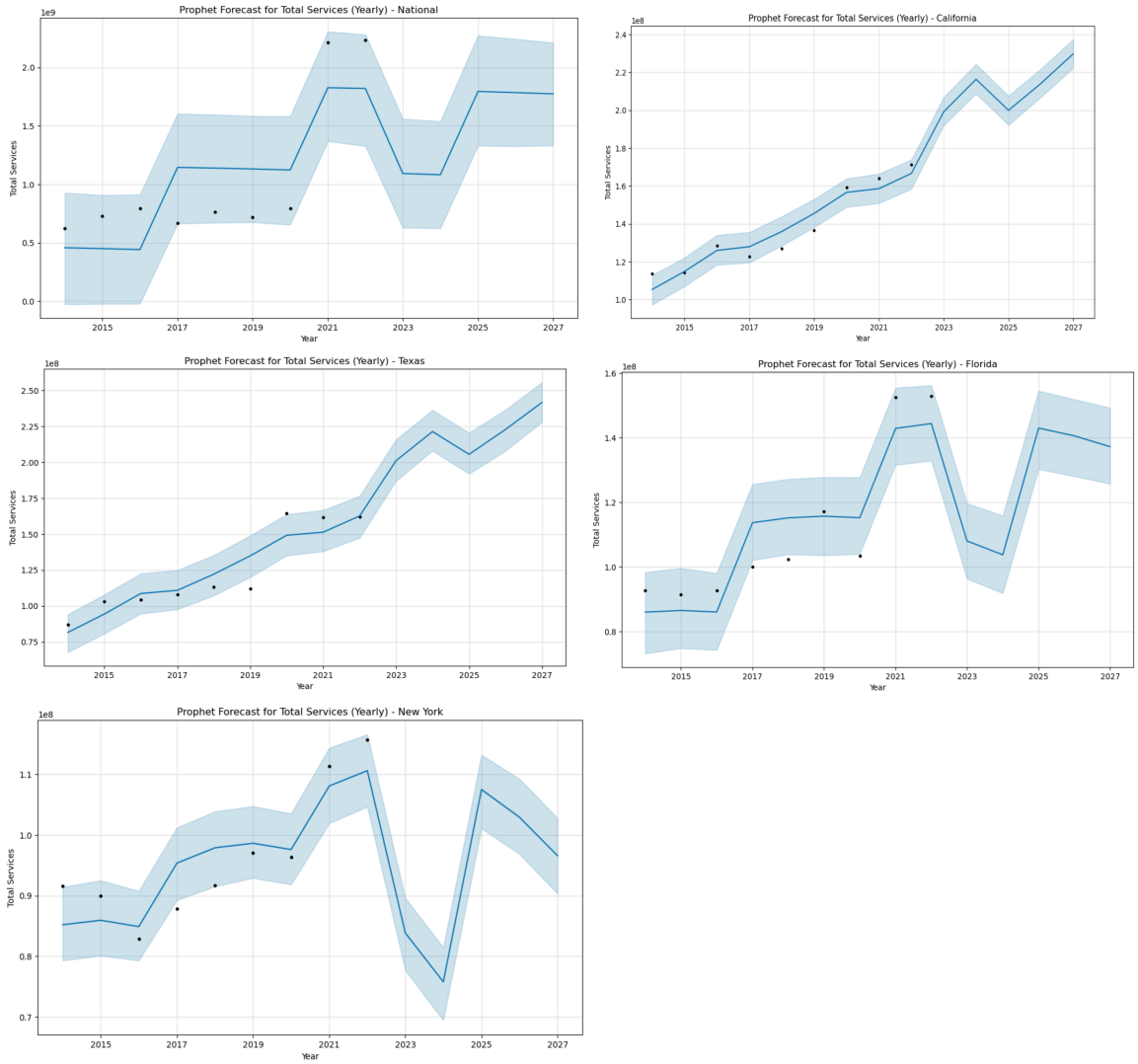
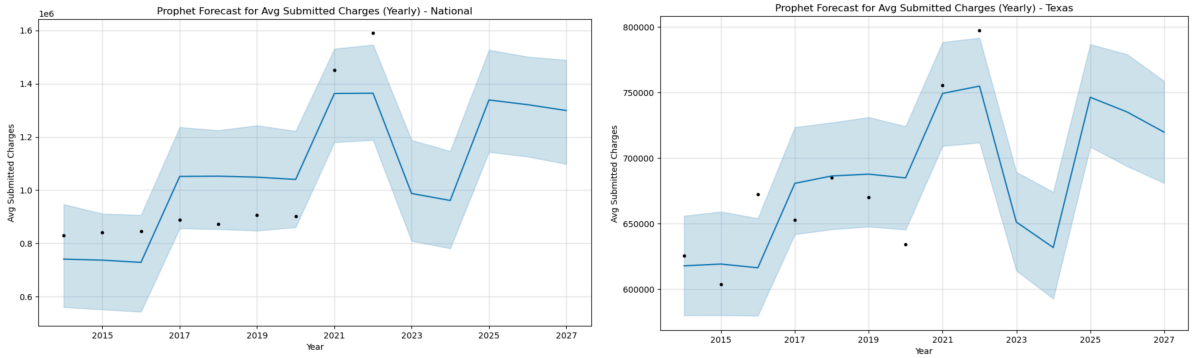


Figure 17 : Forecast Total Services for the Top 5 Regions



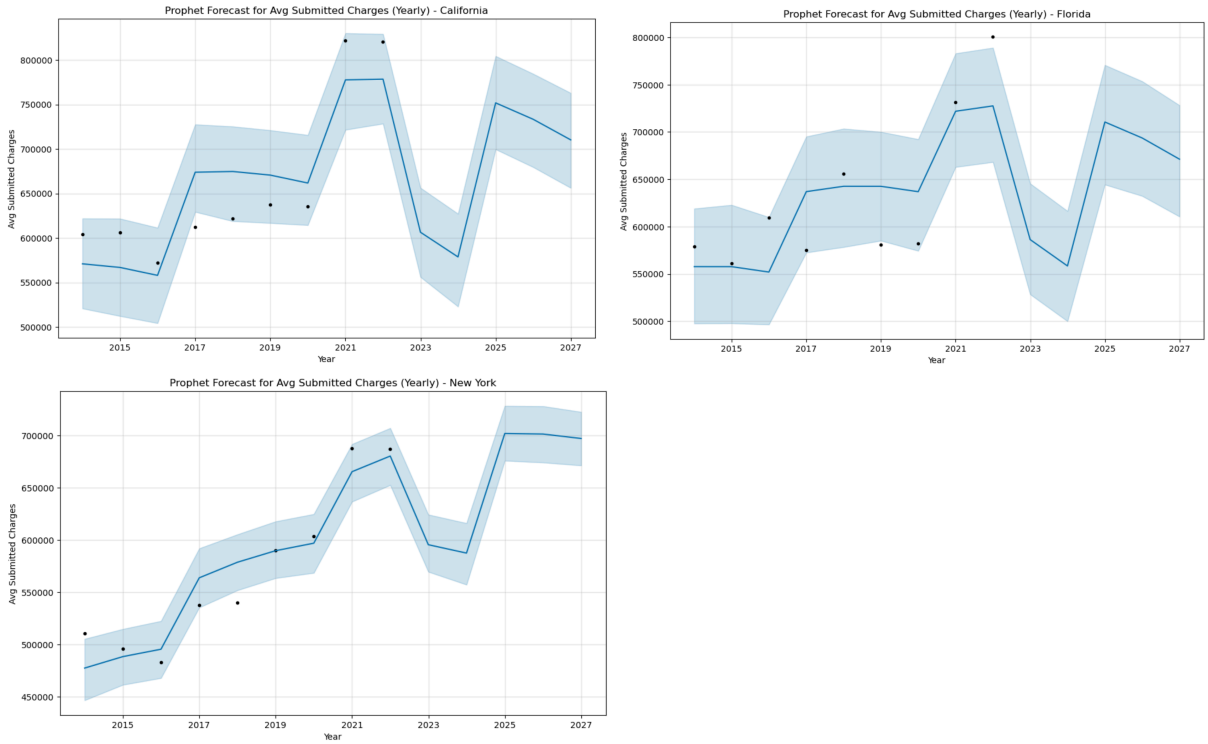


Figure 18 : Forecast Avg Submitted Charges for the Top 5 Regions

Region	Common Trend	Key Differences Between Models
National	Stabilization at 1.6M post-2022	Prophet highlights greater uncertainty post-2023, while ARIMA projects rigid stability.
Texas	Plateau at around 800K	Prophet captures variability pre-2023, ARIMA suggests a smoother trend.
California	Stabilization at 850K	Similar trends, but Prophet highlights potential future dips.
Florida	Stabilization at 750K	The Prophet shows larger uncertainty intervals; ARIMA forecasts steady trends.

New York	Stabilization at 700K	Both models align closely for New York due to consistent historical patterns.

Table 13 - Key Observations Across Regions

The forecast for total supplier services provides further insights into service volumes by region. Nationally, there was a steep rise in total services between 2018 and 2021, followed by stabilization at approximately 2 billion by 2027. Regions like California and Texas are forecasted to continue growing in service volumes, reflecting increasing demand, while Florida and New York show signs of stabilization or slight decline, suggesting possible saturation or policy impacts. High-growth regions like Texas may require more resources to sustain service levels, while stabilizing regions like Florida could benefit from more focused cost management strategies. These insights provide valuable information for planning future resource allocation and optimizing costs in healthcare services.

The Prophet forecasts for average submitted charges reveal key trends across national and regional levels. Nationally, there was a sharp increase in average submitted charges beginning in 2020, which plateaued post-2022. This trend is mirrored across regions like California, Florida, and New York, where charges increased sharply post-2020 before stabilizing. Texas displayed significant variability prior to 2020 but has since seen charges stabilize around 800,000. These trends suggest that, while charges have sharply risen during the pandemic years, they are expected to remain steady over the next few years. Notably, regional differences, such as high

peaks in the national trend and steady growth in Texas, indicate areas requiring targeted intervention and further analysis to address cost drivers and manage future expenses.

In summary, the top 5 regions—California, Texas, Florida, New York, and National—show varying trends in supplier charges and total services. California and Texas stand out as regions exhibiting consistent growth in both average submitted charges and total supplier services, indicating a strong demand and the need for resource allocation to support this growth. In contrast, Florida and New York show signs of stabilization or decline in either charges or services, suggesting a more saturated or controlled healthcare environment where cost optimization may be prioritized. The forecasts emphasize the importance of tailored resource allocation and cost management strategies to address the unique needs of each region.

The evaluation of regional trends in supplier charges reveals that specific regions have experienced consistent growth in charges post-2018. Identifying these temporal patterns allows for proactive resource allocation by highlighting areas with inefficiencies and opportunities for cost optimization. Notable fluctuations in supplier charges during certain periods, particularly in 2020 and 2021, suggest the influence of external factors like the COVID-19 pandemic, which affected healthcare costs significantly.

By understanding these regional variations over time, policymakers can better allocate funds to underserved areas, mitigate inefficiencies, and optimize overall Medicare expenses. This

approach emphasizes the importance of targeted interventions to manage regional disparities and improve cost-effectiveness in healthcare service delivery.

Policy Recommendations for RQ4

1. **Targeted Resource Allocation**: Regions experiencing consistent growth, like California and Texas, should receive additional resources to meet increasing service demand. Proactive planning is necessary to ensure these regions can maintain service quality while managing rising costs.
2. **Cost Management Strategies for Stabilizing Regions**: Regions such as Florida and New York, which are showing signs of stabilization or decline in service growth, should focus on cost management strategies to optimize healthcare expenses without compromising service delivery.
3. **Addressing Inefficiencies**: The sharp increases in supplier charges during the pandemic period highlight the need for targeted interventions to address inefficiencies and manage rising costs. Policymakers should focus on understanding the drivers of these cost spikes and implement policies to mitigate future risks.
4. **Benchmarking Best Practices**: Regions like New York, which show effective cost control and stable trends, can serve as benchmarks for other regions. Best practices from these areas can be applied to improve cost-effectiveness across regions experiencing higher volatility or inefficiencies.

Limitations

1. **Failure to Meet Acceptance Criteria**: Both ARIMA and Prophet models failed to meet the predefined RMSE and MAE thresholds of ≤ 10%. This limitation highlights that the models are not suitable for precise prediction in this context.
2. **Overfitting in ARIMA**: ARIMA demonstrated lower errors on training data compared to test data, indicating a potential overfitting issue, which reduces its reliability for generalization.
3. **Consistency in Prophet**: While Prophet showed consistent performance between training and test data, its errors were still above acceptable levels, especially in the presence of anomalies, suggesting insufficient adaptability to data irregularities.
4. **Anomalies and Non-Linear Trends**: Both models struggled with capturing non-linear trends and handling anomalies in the data, which significantly impacted their forecasting accuracy.
5. **Limited Data Transformation and Hyperparameter Tuning**: Attempts to perform log transformation or hyperparameter tuning for both models were unsuccessful due to system crashes or extremely long computation times. Despite indications from residuals and QQ plots that log transformation could improve performance, these transformations were not feasible. This restricted the ability to optimize model performance fully.
6. **Weak Stationarity**: The dataset exhibited weak stationarity, which complicated the forecasting process. Both models struggled to adequately handle this, impacting their ability to make accurate predictions over time.

Future Work

1. **Explore Advanced Deep Learning Models**: Consider utilizing advanced time-series models, such as LSTM or GRU, which may better capture complex non-linear trends and reduce overfitting.
2. **Hyperparameter Tuning and Feature Engineering**: Conduct extensive hyperparameter tuning and feature engineering to enhance model performance, including methods that can adapt to anomaly-driven variations.
3. **Combining Multiple Models**: Use a hybrid approach by combining ARIMA, Prophet, and other machine learning models to leverage their individual strengths for better overall accuracy.
4. **Addressing Computational Limitations**: Implement more efficient computational methods or use more powerful hardware to enable transformations like log scaling and hyperparameter tuning, which showed potential for improving model accuracy in initial analyses.
5. **Improving Stationarity**: Apply advanced differencing or transformations to achieve stronger stationarity in the dataset, thereby improving the suitability of models like ARIMA and Prophet for accurate forecasting.

Discussion

This study demonstrates the power of a multi-method approach in identifying and addressing inefficiencies in healthcare supplier practices. By leveraging anomaly detection, classification, correlation analysis, and forecasting, it provides actionable insights for policymakers to optimize resource allocation, improve equity, and enhance the efficiency of healthcare delivery. While challenges persist in modeling complexities, the results lay the foundation for future research and innovation, driving systemic improvements in the healthcare sector.

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)

10) Centers for Medicare & Medicaid Services (CMS). (n.d.). Geographic variation in standardized Medicare spending by state. https://data.cms.gov/tools/geographic-variation-in-standardized-medicare-spending-stat

API

- Scikit-Learn. (n.d.). *sklearn.ensemble.IsolationForest*. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.IsolationForest.html
- Analytics Vidhya. (2021, July). *Anomaly Detection using Isolation Forest: A Complete Guide*.https://www.analyticsvidhya.com/blog/2021/07/anomaly-detection-using-isolation-forest-a-complete-guide/
- Scikit-Learn. (n.d.). *sklearn.model_selection.cross_val_score*. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_val_score.html

- Scikit-Learn. (n.d.). *sklearn.ensemble.RandomForestClassifier*. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Books

1. Data Preparation for Machine Learning - Data Cleaning, Feature Selection, and Data Transforms in Python (Jason Brownlee)

Appendix A

Table 14 – 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* A concatenation of the RBCS category and subcategory description	Nominal (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*	Discrete (Quantitative)

	The total number of referring providers ordering DMEPOS products/services	
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* The total number of DMEPOS products/services rendered by suppliers	Discrete (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 15 – 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

Revision History

Date	Description	Reason
2024-11-27	Update title	The project title was revised to improve clarity and specificity
2024-11-27	Removed RQ5 and consolidated insights	RQ5 was redundant; covered by RQ1, RQ2, RQ3 and RQ4.
2024-11-27	Refined RQ1 and RQ2 to clarify intended outcomes	Improved specificity by connecting RQ1 to Medicare resource allocation and RQ2 to actionable policy recommendations.
2024-11-27	Expanded literature review to include recent CMS statistics on DME spending disparities.	Provided concrete data to highlight gaps in Medicare's resource allocation and justify the study's objectives
2024-11-27	Expanded the Introduction section to discuss the underexplored nature of DMEPOS resource allocation and its connection to fraud detection methodologies.	1) Provided concrete data to highlight gaps in Medicare's resource allocation and justify the study's objectives 2) Explained how fraud detection insights inform the study's approach to identifying inefficiencies.
2024-11-27	Updated the underserved_level variable to use an "or" condition for classification.	Ensured sufficient variability in the data to support robust model development.

2024-11-27	Updated the abstract to highlight the study's unique focus on geographic disparities, inefficiencies, and policy impacts	Highlighted the study's focus on underserved regions, temporal trends, and the application of machine learning and time-series analysis.
2024-11-27	Removed the Apriori section from the methodology.	PCA analysis demonstrated that categorical variables were not significant, making Apriori unnecessary.
2024-11-27	Insights from Spearman’s correlation were documented to identify variables contributing to inefficiencies.	To explain how the analysis supports model building and feature selection.
2024-11-27	Updated the methodology section to include revised details on anomaly detection, classification, and time-series forecasting. Added evaluation metrics such as RMSE, MAE, and MSLP.	Refined methodology based on notebook validations and clarified the performance metrics used in the analysis.