UNSUPERVISED ANAMOLY DETECTION

Healthcare Providers Data For Anomaly Detection

PROJECT PRESENTATION

BY SHRIKAR GAIKAR

FLOW OF PRESENTATION

- Problem Statement
- Dataset
- Exploratory Data Analysis
- Preprocessing
- Clustering
- ML Algorithms
- SHAP Analysis
- DL Algorithm

PROBLEM STATEMENT

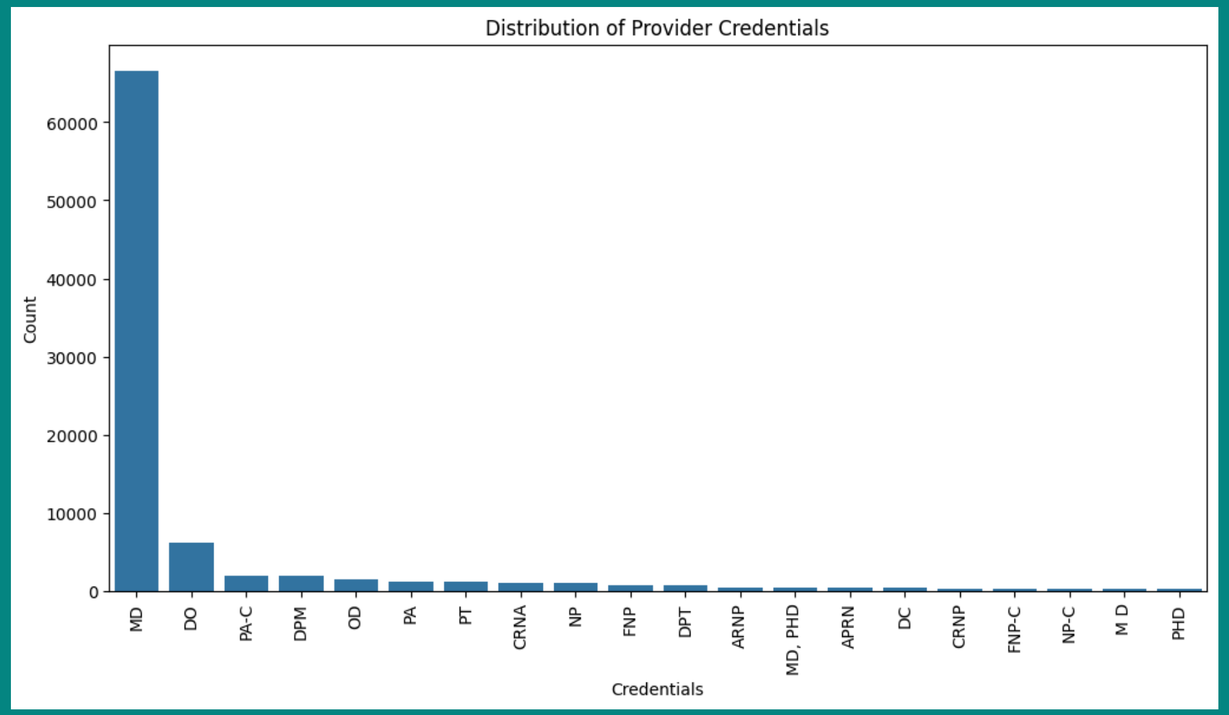
Detecting Healthcare Fraud Using Data Analysis

- Initially, analyzing healthcare data to detect fraud may seem daunting due to its complexity.
- However, by systematically examining patterns and anomalies in the data, we can uncover potential instances of fraud.
- Our goal is to use data-driven methods to identify irregularities and prevent fraudulent activities within healthcare services.
- This approach ensures transparency and integrity in healthcare practices, safeguarding against misuse and ensuring fair treatment for all patients and providers.

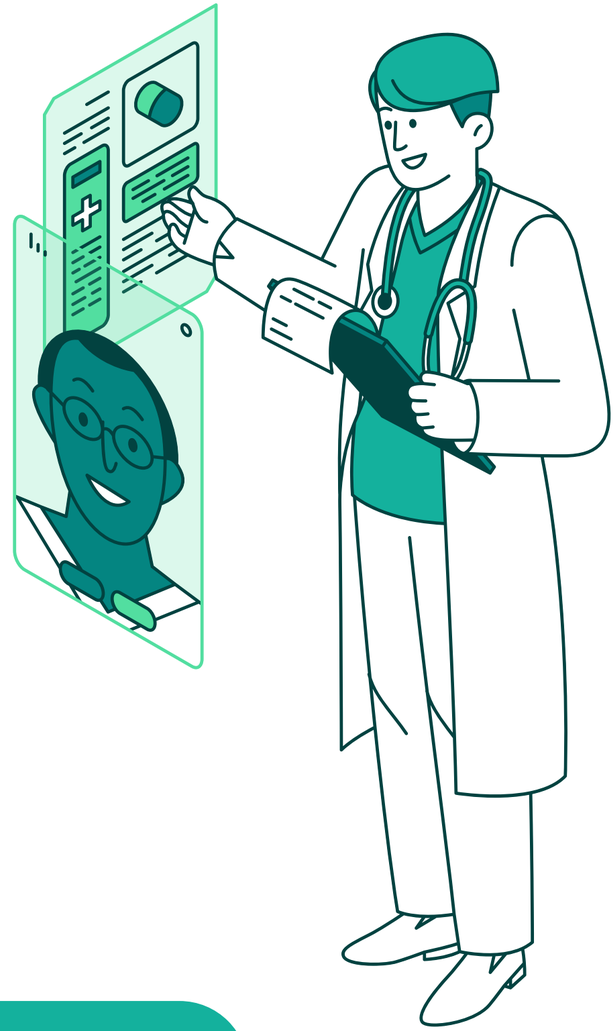
DATASET

- The dataset used for this project contains detailed information about healthcare providers, including various metrics related to their services and payments.
- It includes data such as **National Provider Identifier (NPI), provider names, addresses, credentials, gender, entity type, provider types, service counts, Medicare payment amounts,** and more.
- This comprehensive dataset allows us to analyze provider behavior, detect anomalies that may indicate potential fraud, and ensure compliance with healthcare regulations.

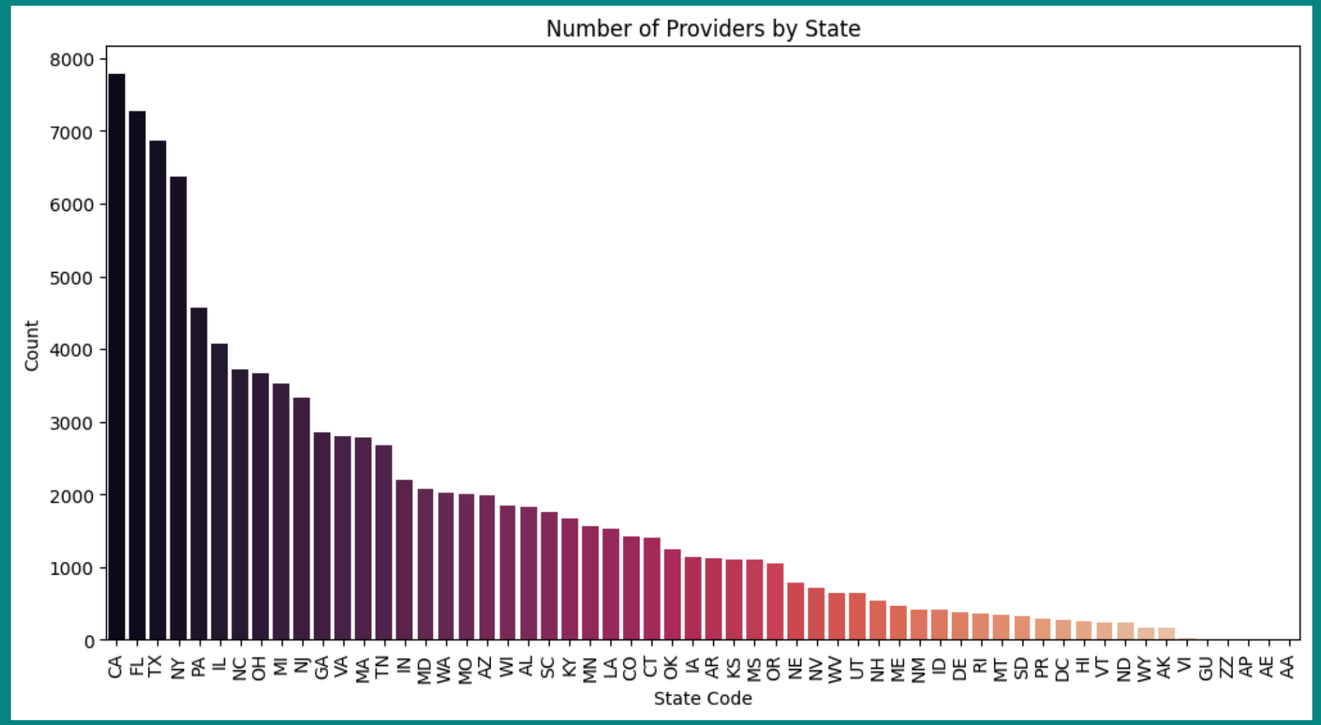
EXPLORATORY DATA ANALYSIS



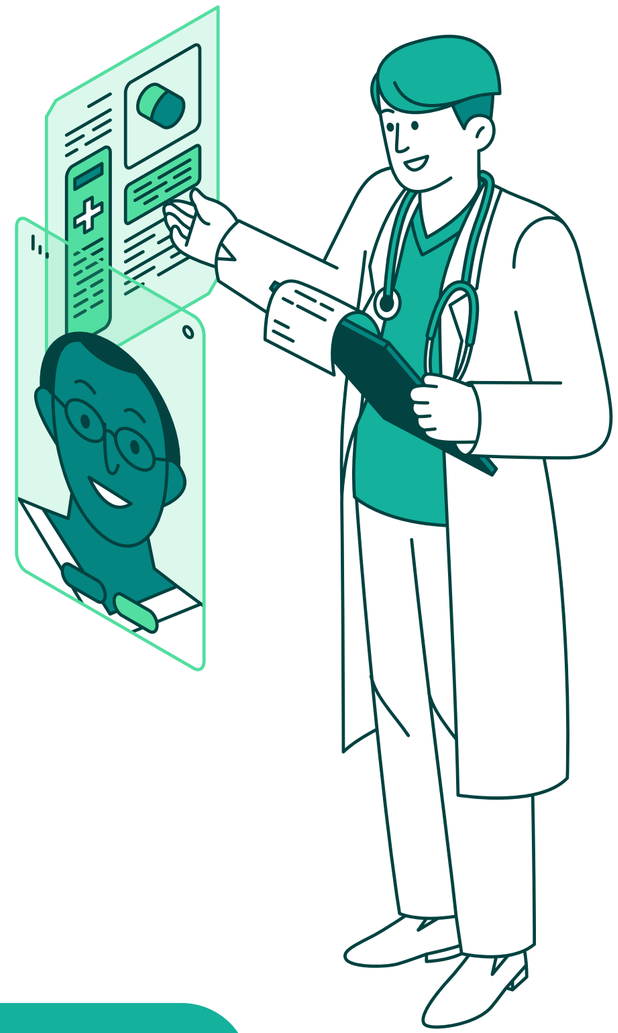
The distribution of provider credentials shows the most common qualifications among healthcare providers in our dataset



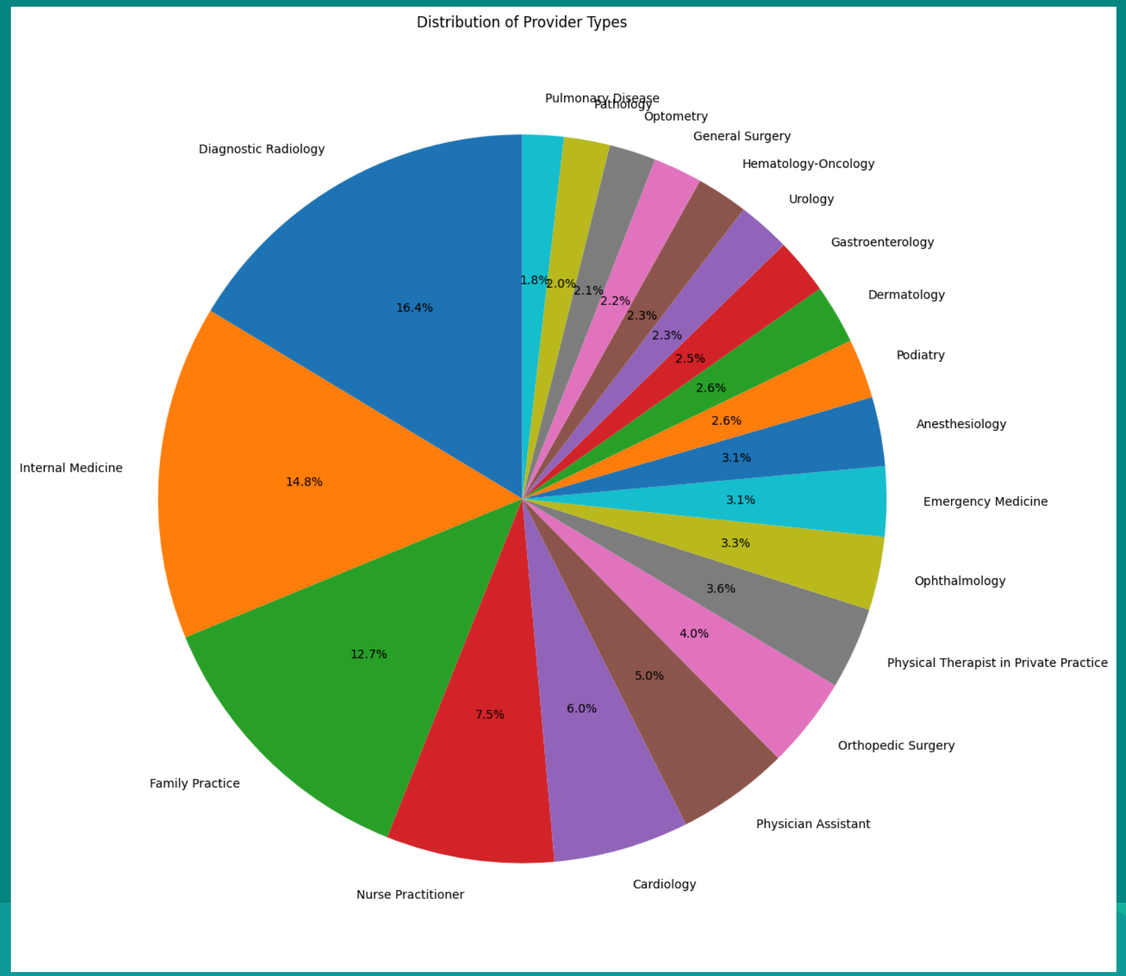
EXPLORATORY DATA ANALYSIS



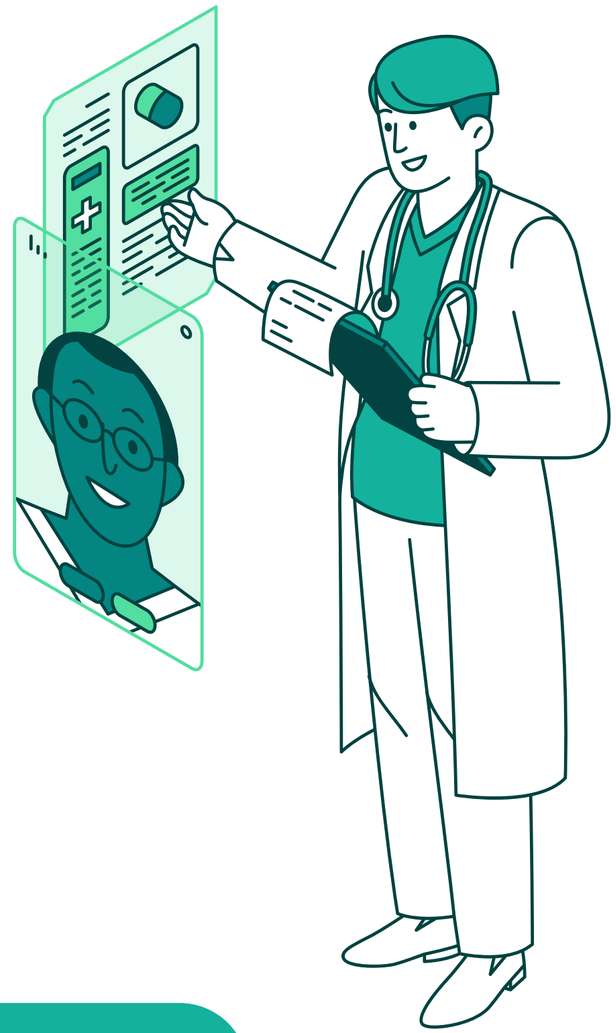
California, Florida and Texas have the highest number of healthcare providers in our dataset.



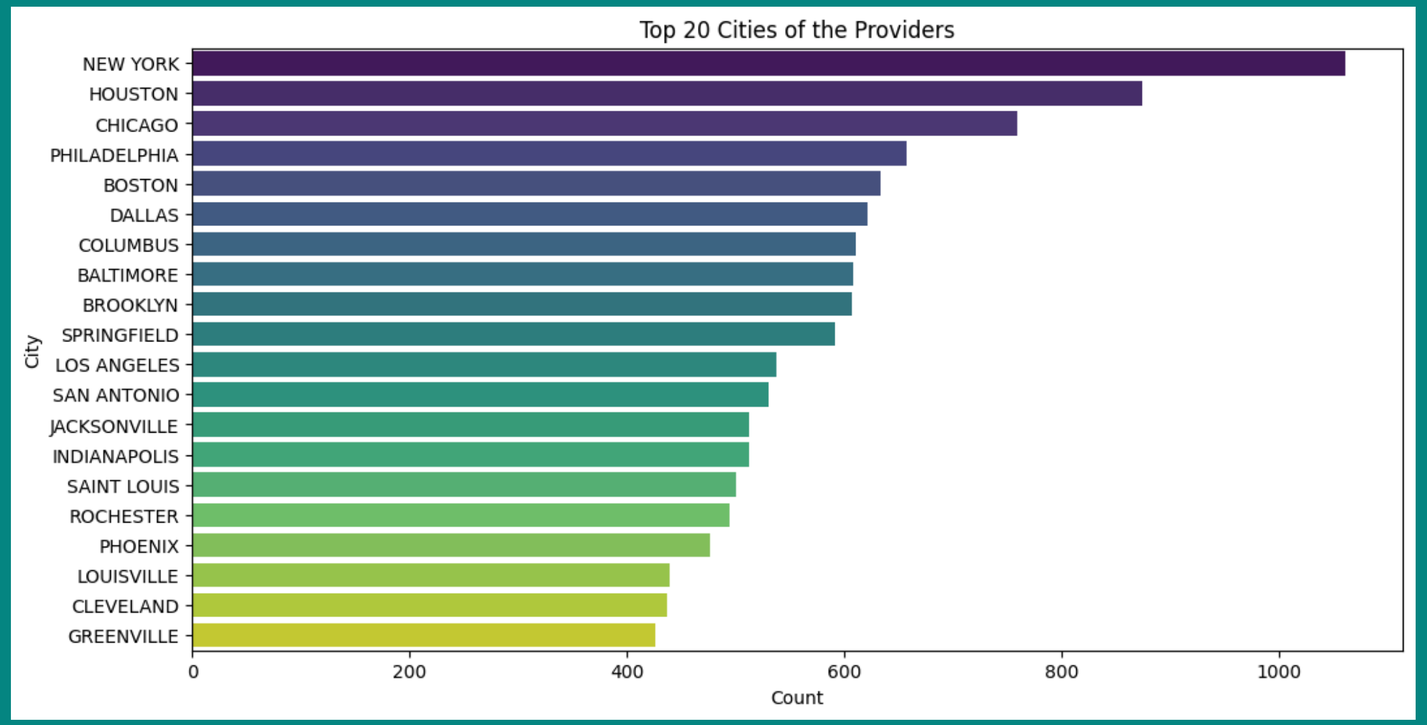
EXPLORATORY DATA ANALYSIS



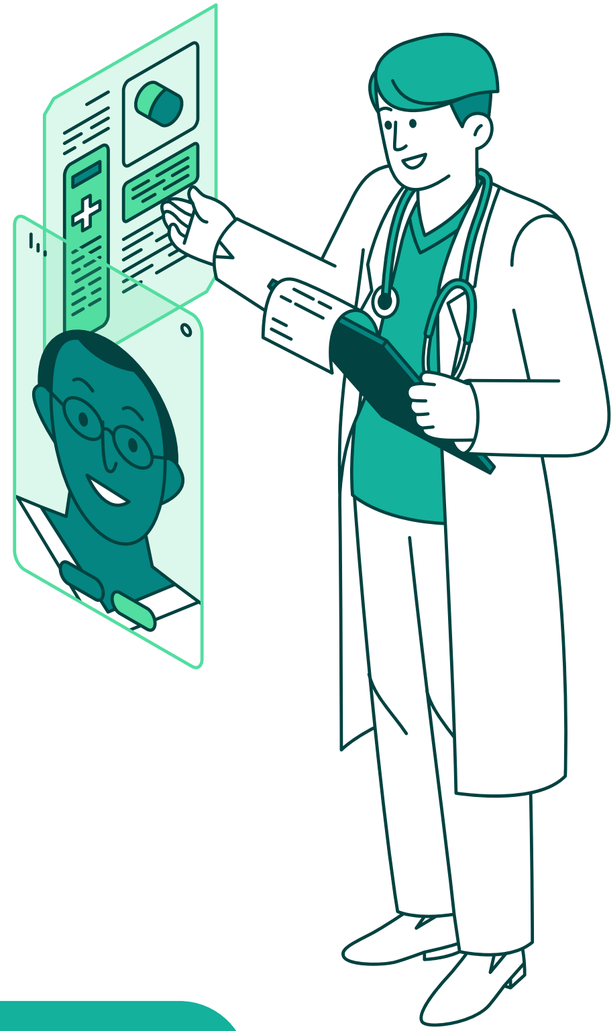
Diagnostic Radiology and internal medicine are the most prevalent provider types in our dataset.



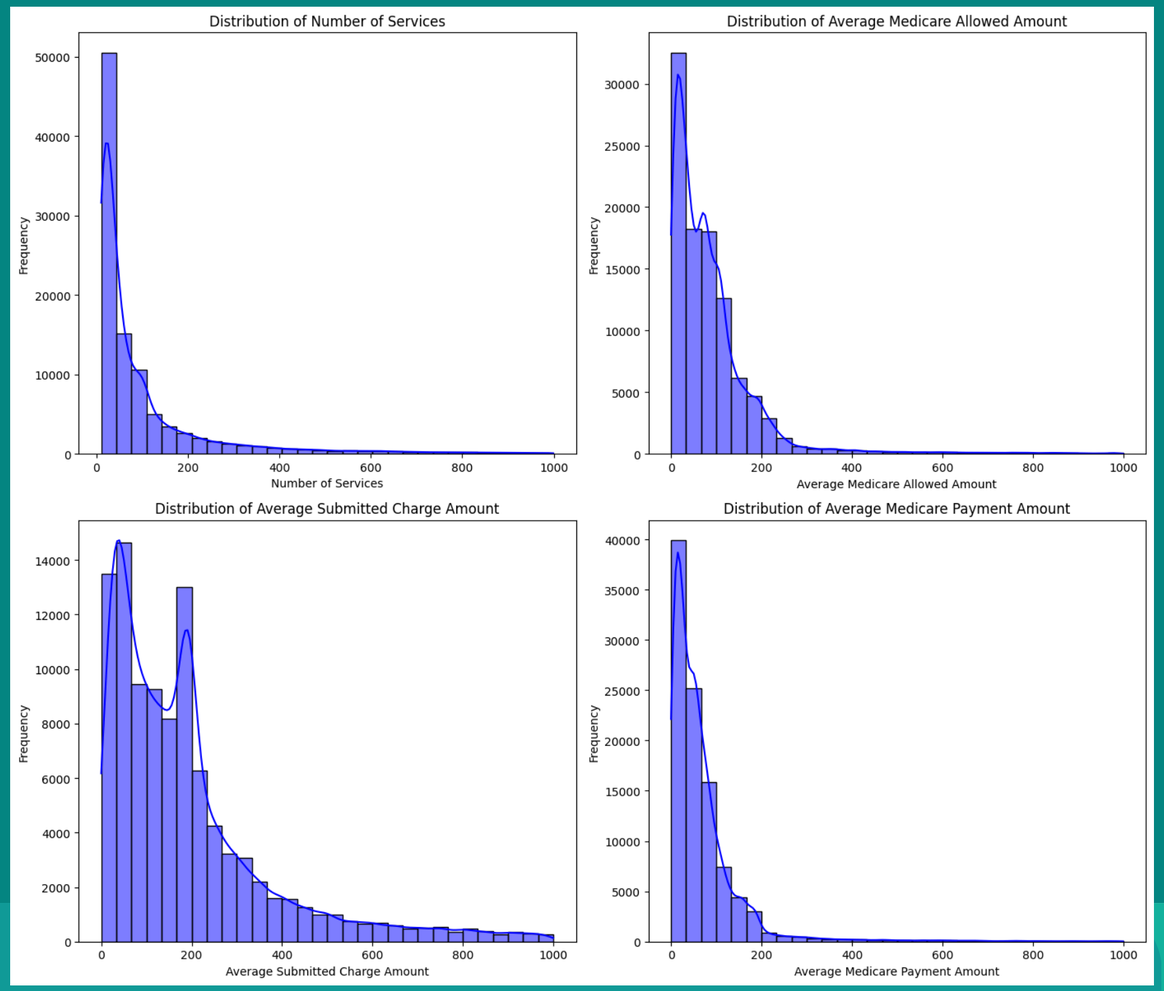
EXPLORATORY DATA ANALYSIS



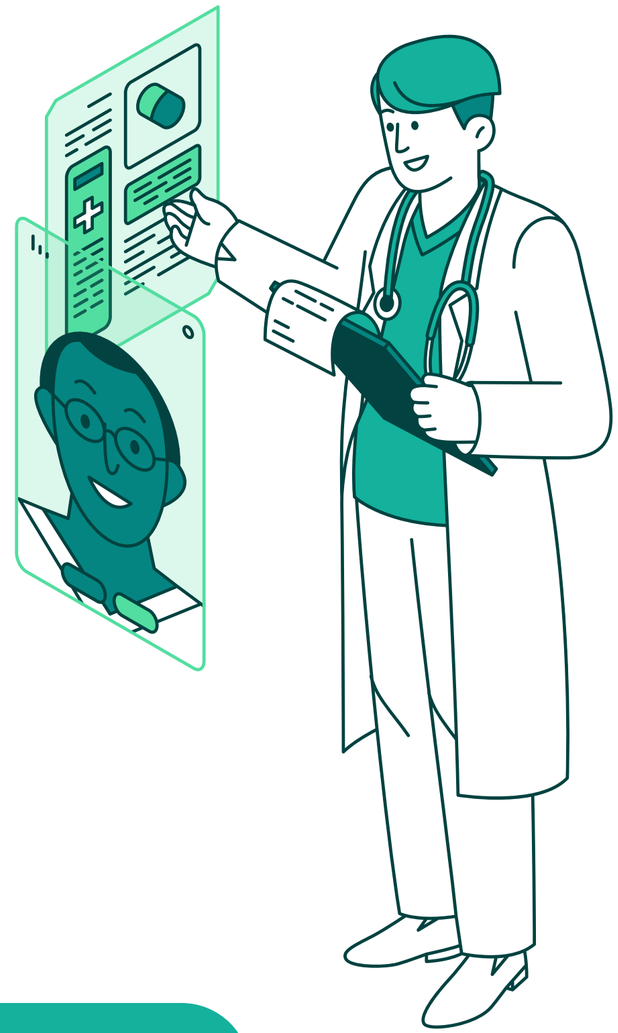
New York City and Houston are among the cities with the highest number of healthcare providers.



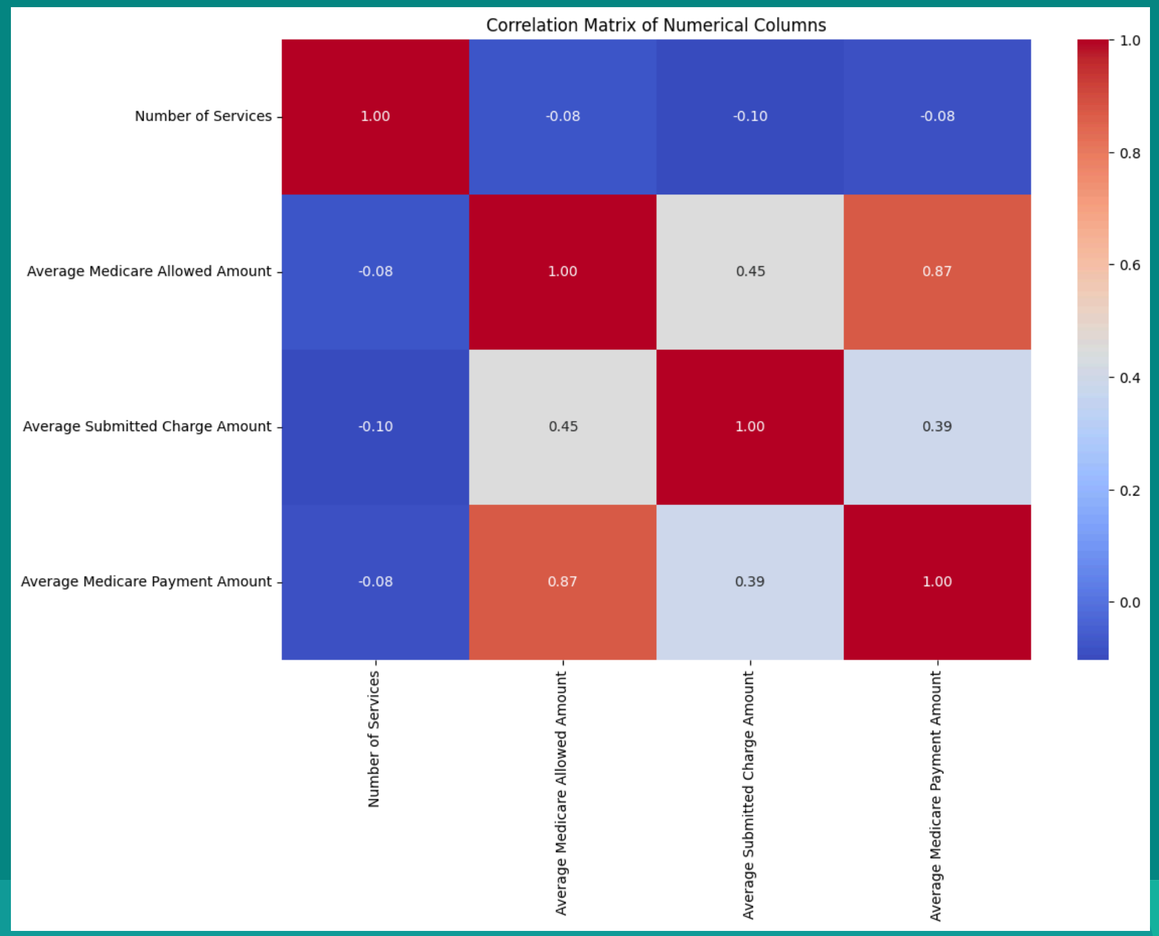
EXPLORATORY DATA ANALYSIS



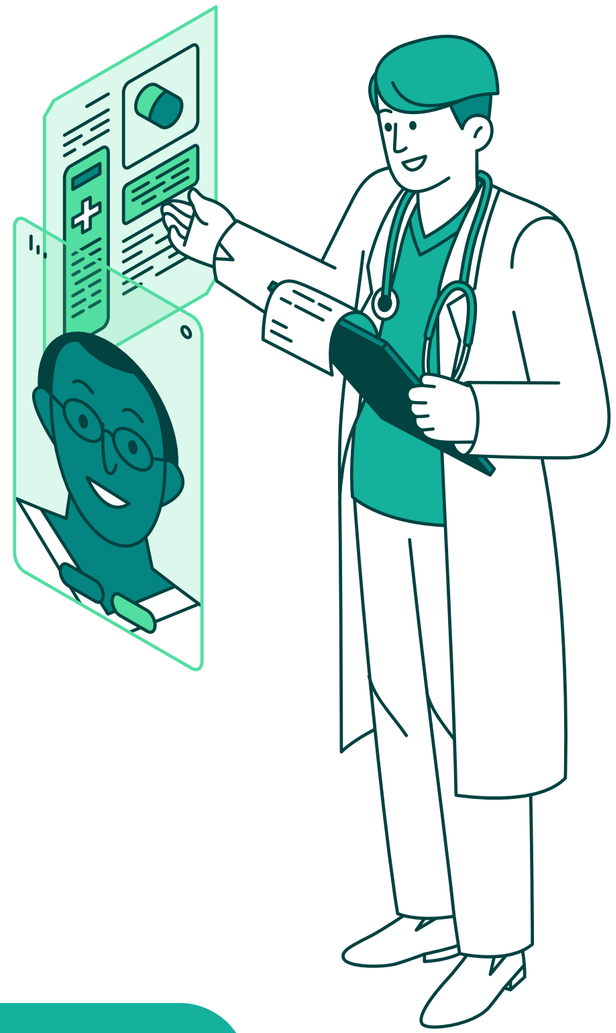
Numeric columns such as number of services and payment amounts show skewed distributions.



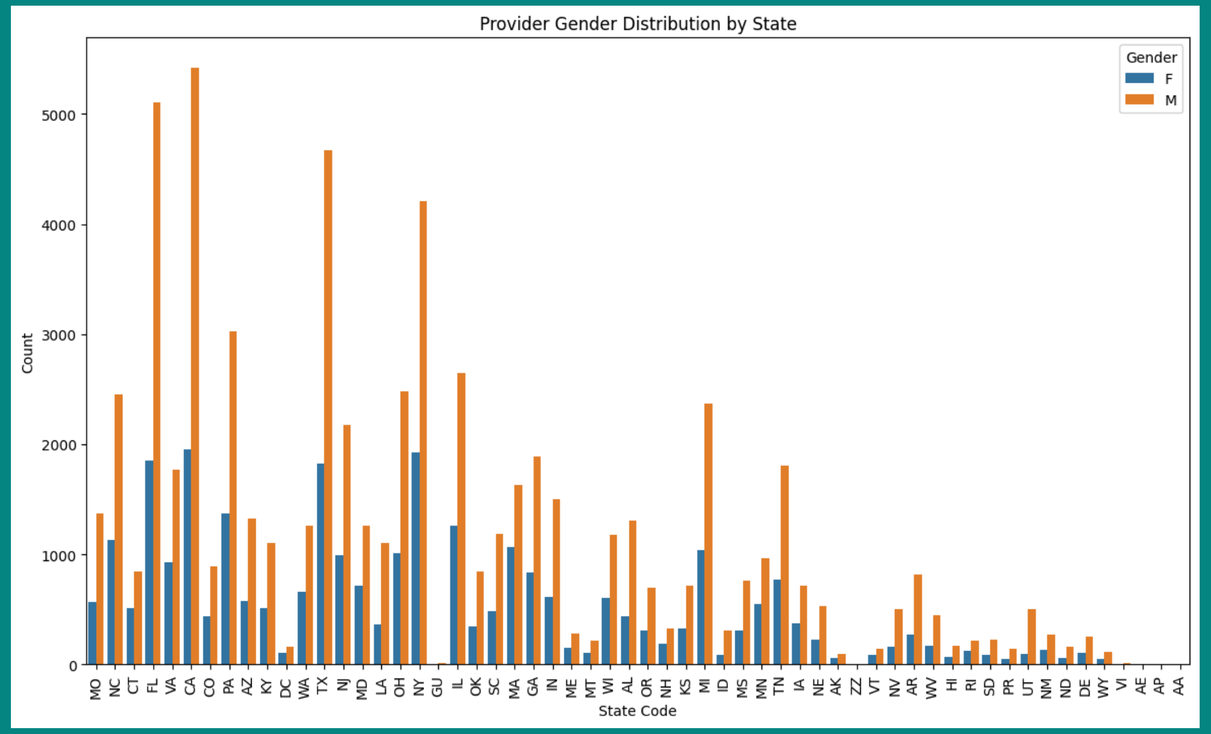
EXPLORATORY DATA ANALYSIS



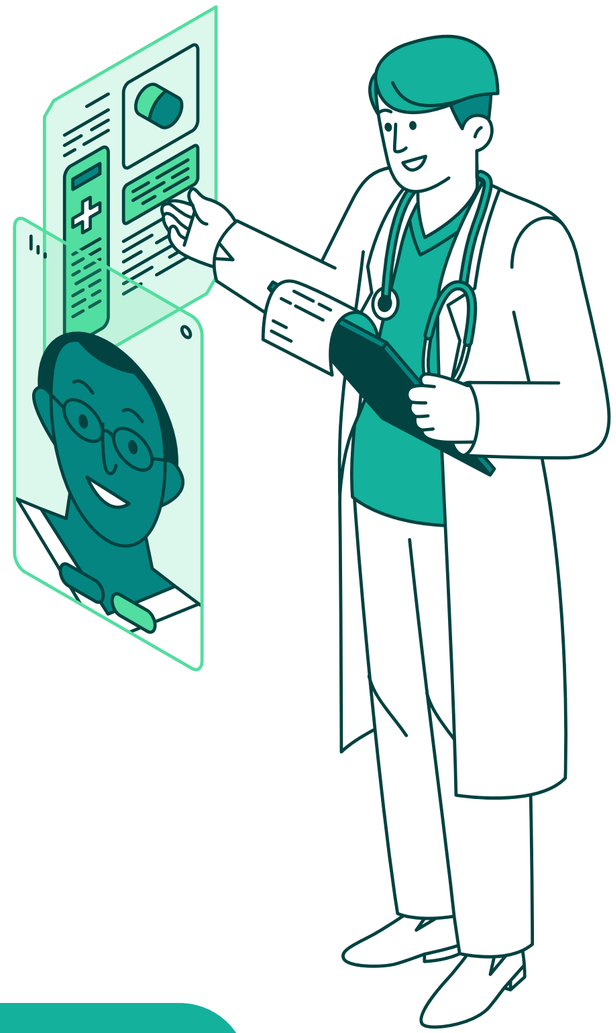
There is a moderate positive correlation between average Medicare payment amounts and other numeric features.



EXPLORATORY DATA ANALYSIS



Gender distribution varies by state, with some states showing more balanced distributions than others.



PREPROCESSING

- **Handling Missing Values:** Impute missing values with the mean.
- **Converting Object Columns to Numeric Types:** Convert applicable object-type columns to numeric types.
- **Checking for Duplicate Values:** Remove duplicate entries to ensure data integrity.
- **Merging Name Columns:** Combine first and last names into a single column.
- **Merging Address Columns:** Merge address components into a single address column.
- **Standardizing Credentials Column:** Ensure consistency in the credentials column by applying standard nomenclature.
- **Encoding Categorical Variables:** Convert categorical variables into numerical values using binary or frequency encoding.

CLUSTERING

2 Approaches:

K MEANS

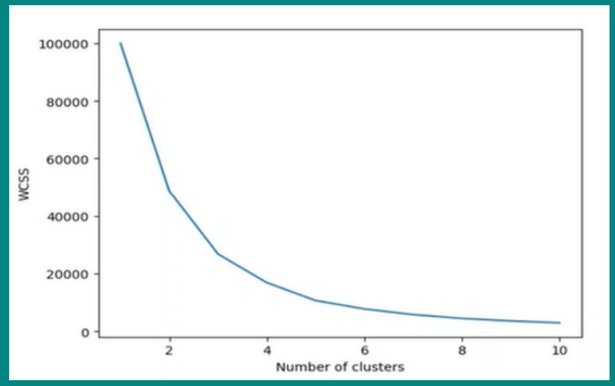
- Groups data into clusters based on similarity.
- Partitions data into a specified number of clusters by minimizing the variance within each cluster. Each data point is assigned to the nearest cluster center.

DBSCAN

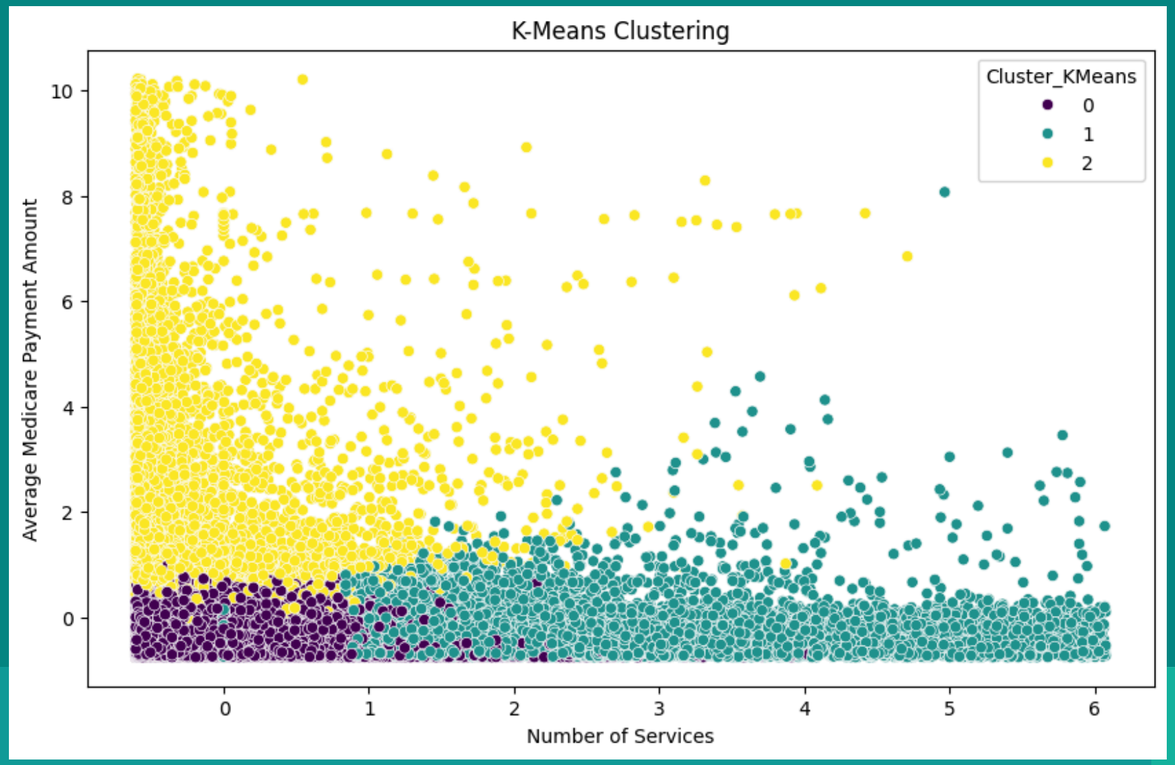
- Groups data points based on density.
- Identifies clusters of high density and marks points in low-density regions as noise. It's effective for finding clusters of varying shapes and sizes.

CLUSTERING

K MEANS

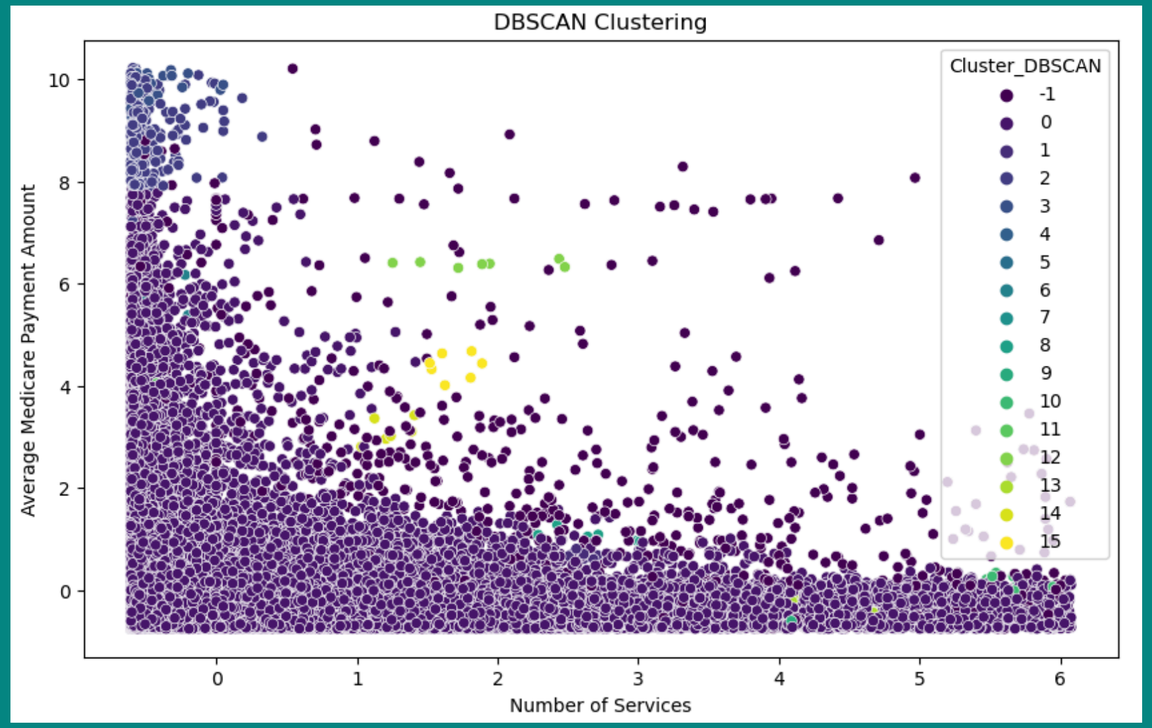


Optimal clusters: 3

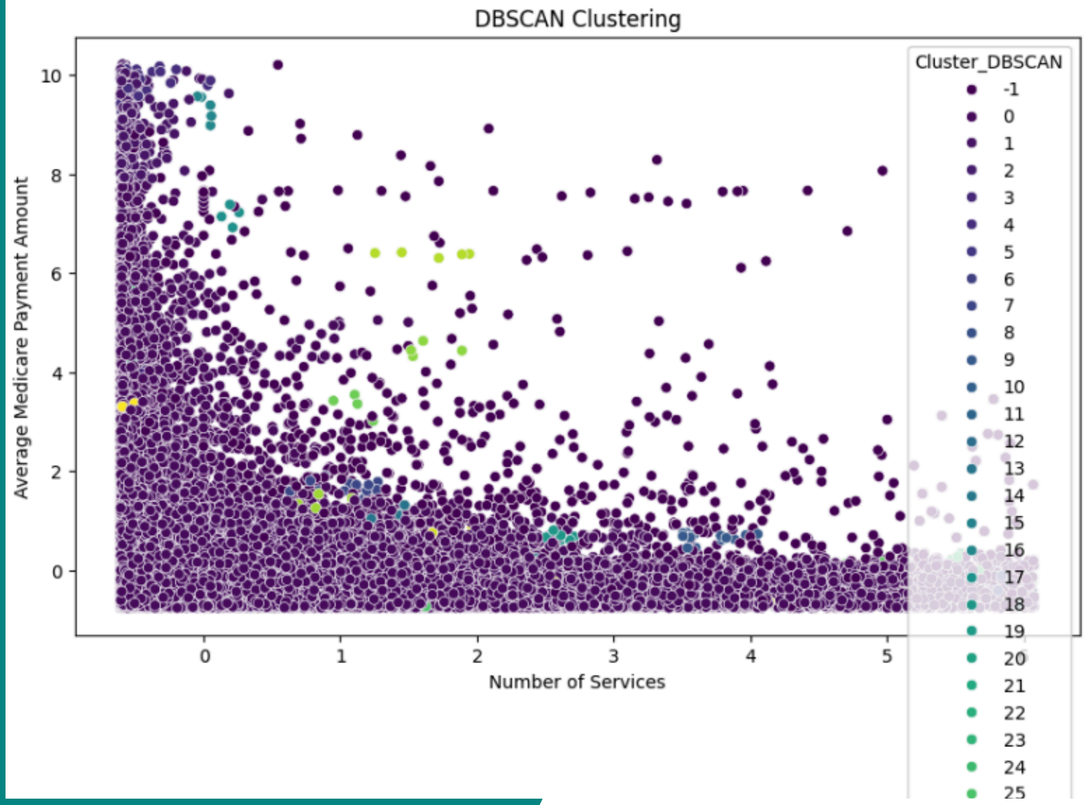


CLUSTERING

DBSCAN



Epsilon radius: 0.7
Number of samples: 6
Found 17 Clusters (-1 to 15)



Epsilon radius: 0.5
Number of samples: 4
Found 27 Clusters (-1 to 25)

ML ALGORITHMS

3 Approaches:

ISOLATION FOREST

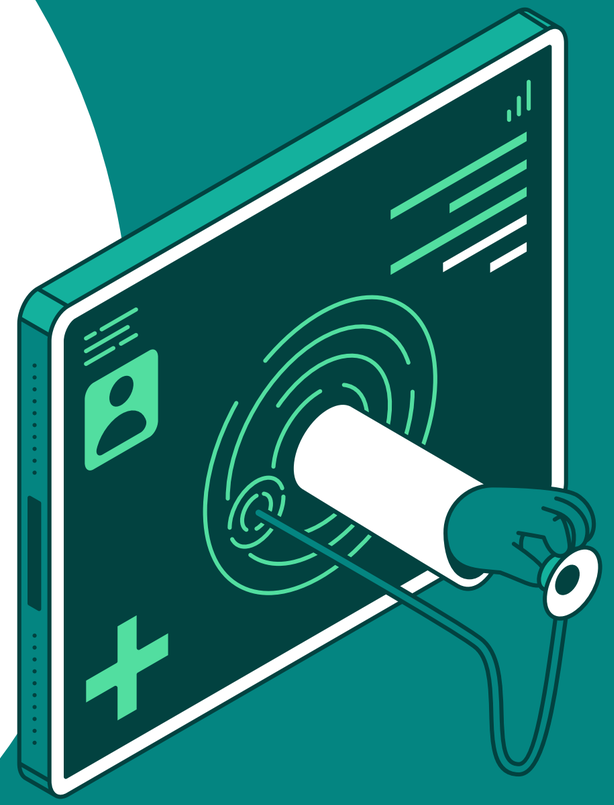
- Identifies anomalies by isolating observations.
- It splits data points repeatedly to see how easily they can be isolated, with anomalies being isolated quickly

ELLIPTIC ENVELOPE

- Detects outliers assuming data follows a Gaussian distribution.
- It fits an ellipse around the data, with points outside the ellipse considered anomalies.

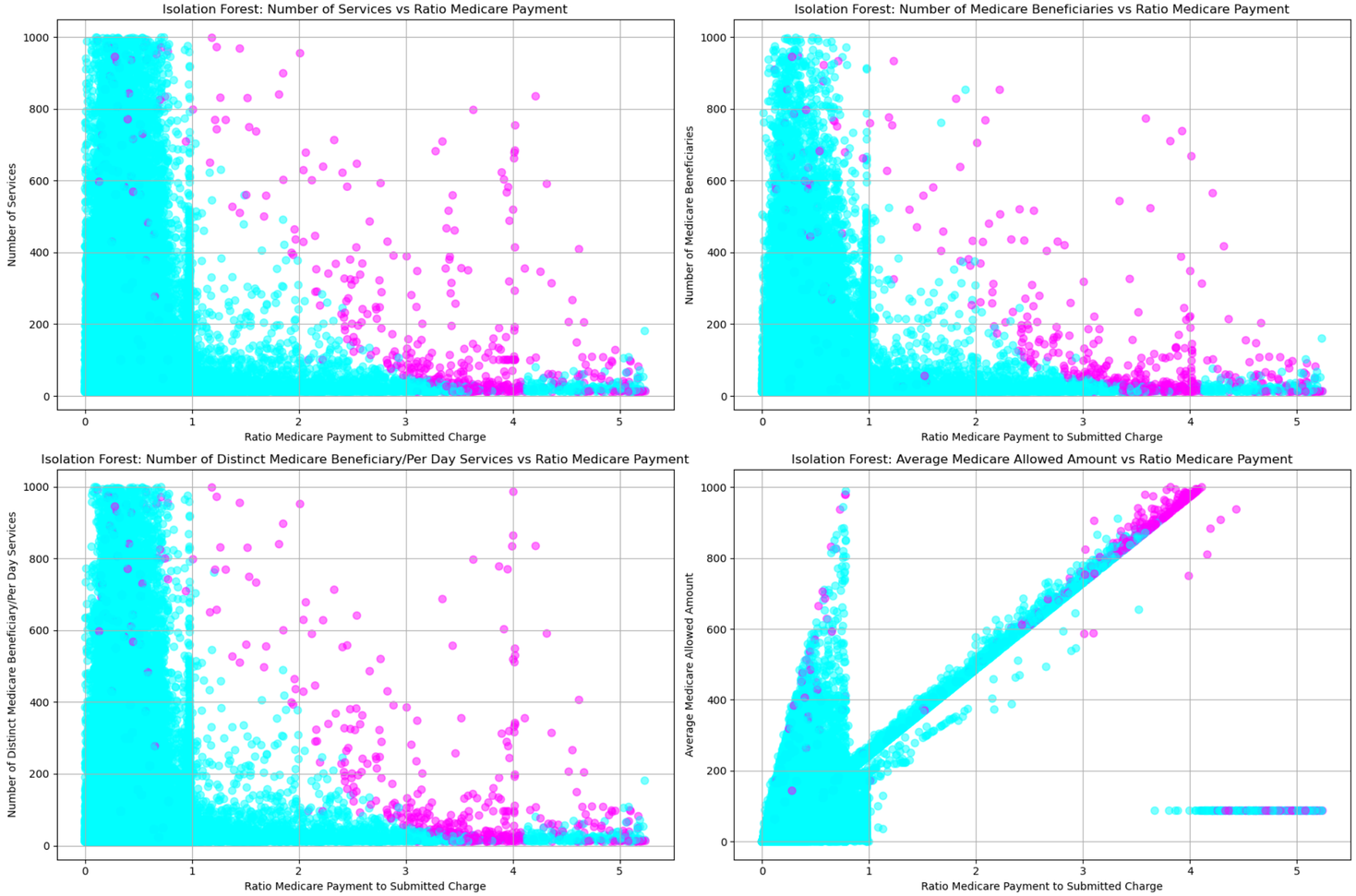
ONE CLASS SVM

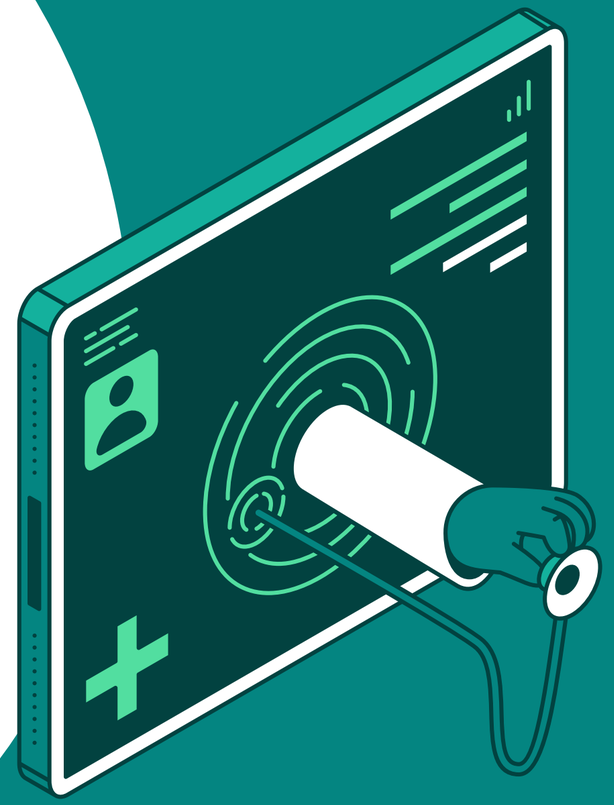
- Classifies data points as normal or outliers.
- It learns the boundary of normal data points in the feature space and identifies points that lie outside this boundary as anomalies.



ML ALGORITHMS

ISOLATION FOREST

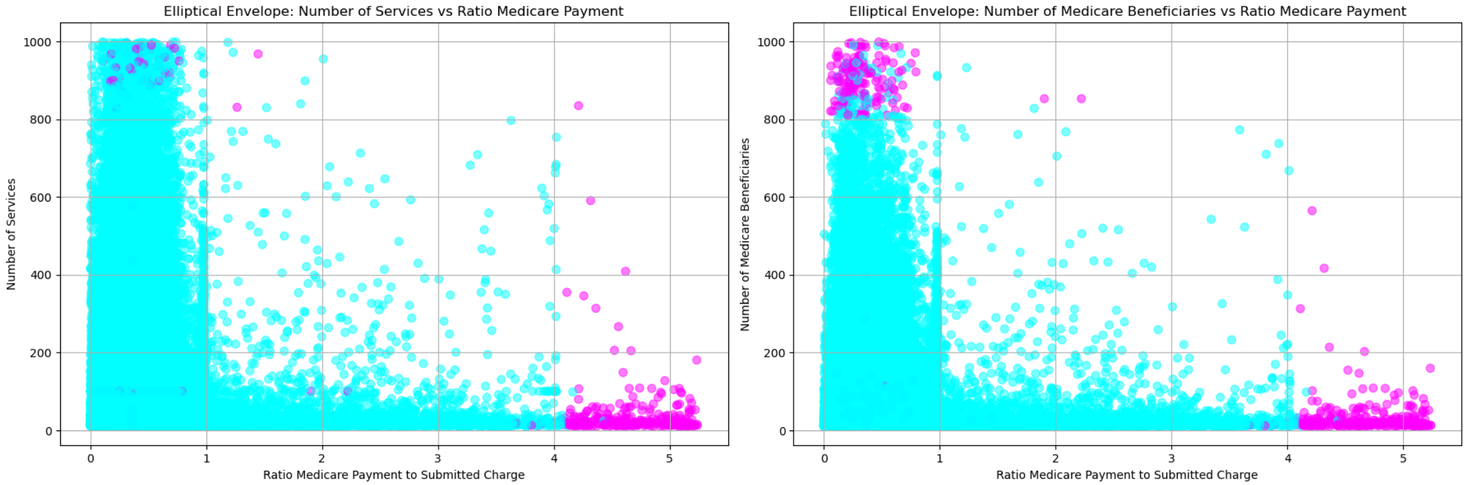


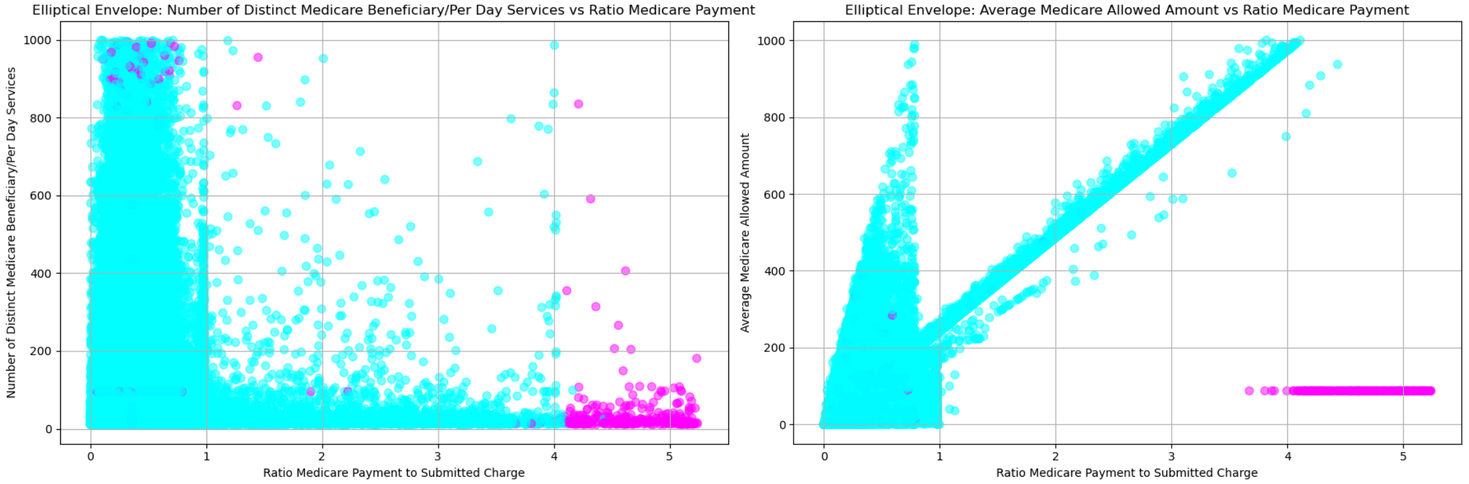


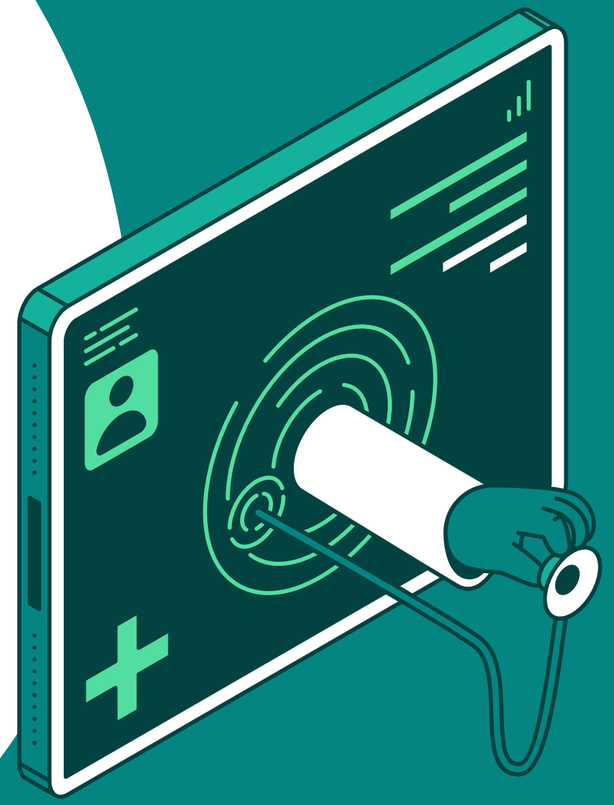
Outliers detected: 500

ML ALGORITHMS

ELLIPTIC ENVELOPE



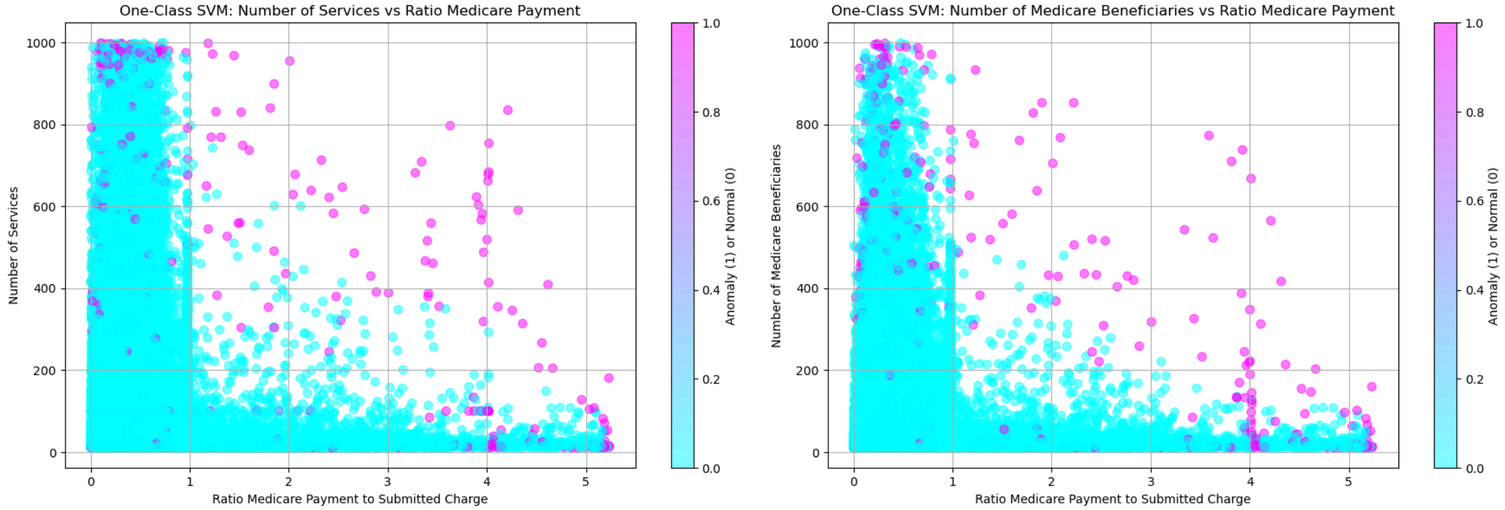


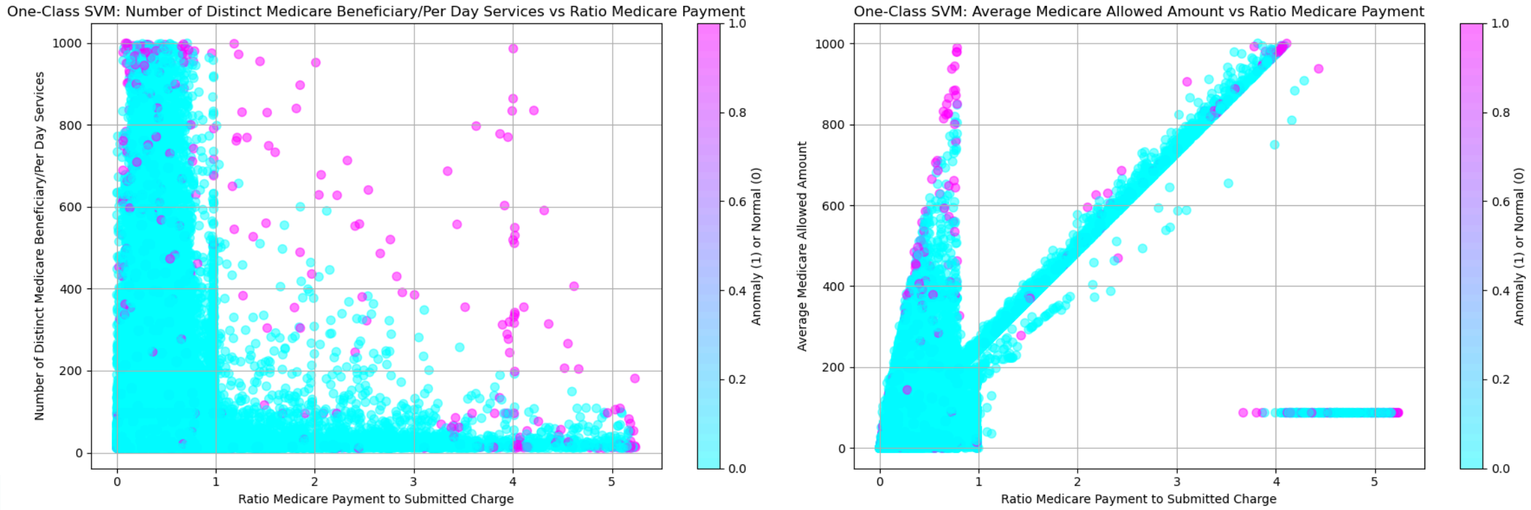


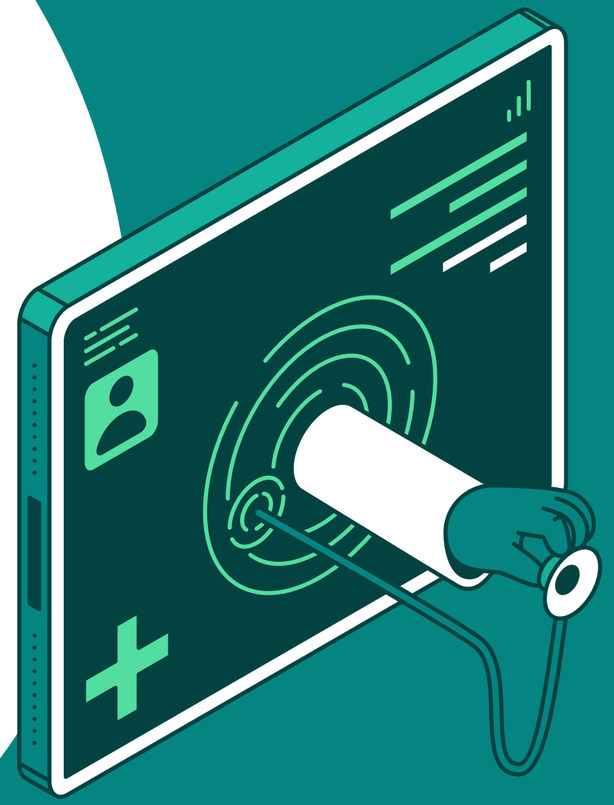
Outliers detected: 500

ML ALGORITHMS

ONE CLASS SVM



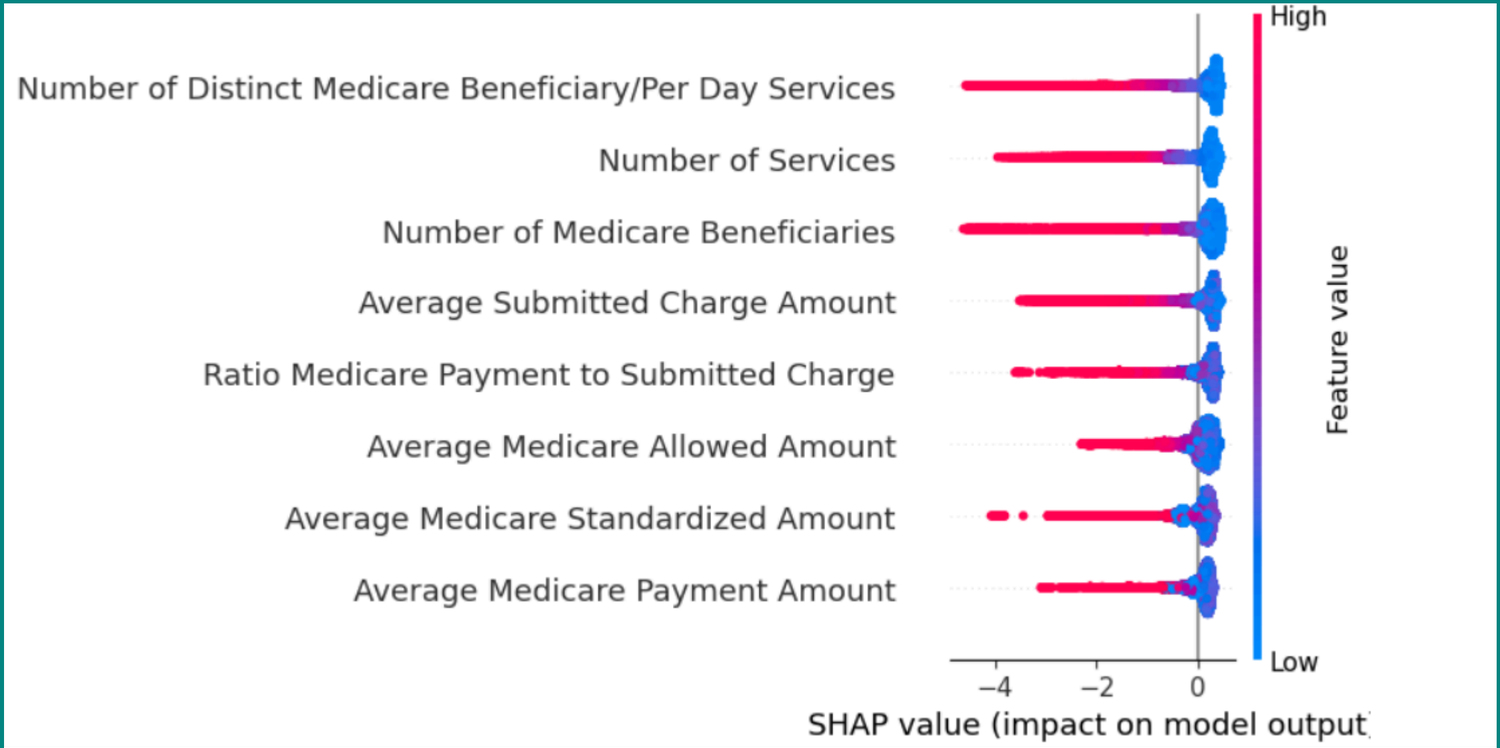




Outliers detected: 505

SHAP ANALYSIS

ISOLATION FOREST



Some columns tend to negatively affect the output

- Number of Distinct Medicare Beneficiary/Per Day Services
- Number of services
- Number of Distinct Medicare Beneficiaries
- Average Medicare Standardized Amount

Inference: Tendency of fraud increases with higher values in these columns

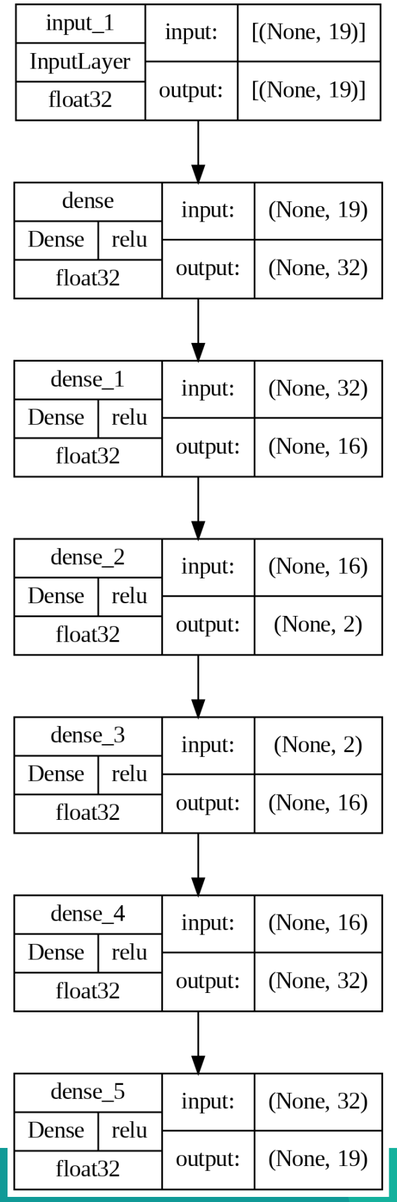
DL ALGORITHM

AUTO ENCODERS

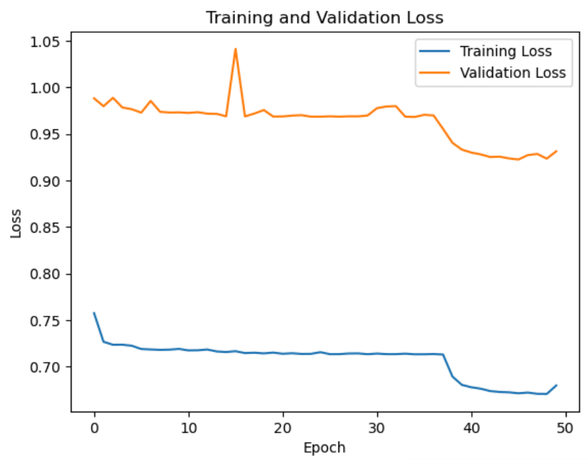
Uses a neural network to detect anomalies by reconstruction error.

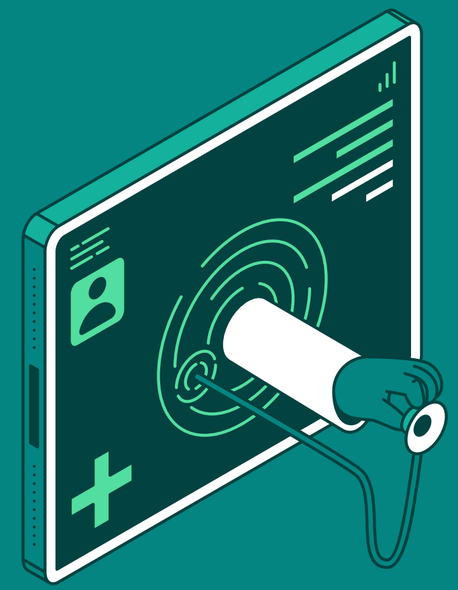
It tries to reconstruct the input data; anomalies are detected when reconstruction error is high because the model cannot accurately reconstruct them.

Outliers detected: 1000




```
Model: "model"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_1 (InputLayer)        [(None, 19)]              0

 dense (Dense)               (None, 32)                640

 dense_1 (Dense)             (None, 16)                528

 dense_2 (Dense)             (None, 2)                 34

 dense_3 (Dense)             (None, 16)                48

 dense_4 (Dense)             (None, 32)                544

 dense_5 (Dense)             (None, 19)                627

=================================================================
Total params: 2421 (9.46 KB)
Trainable params: 2421 (9.46 KB)
Non-trainable params: 0 (0.00 Byte)
```

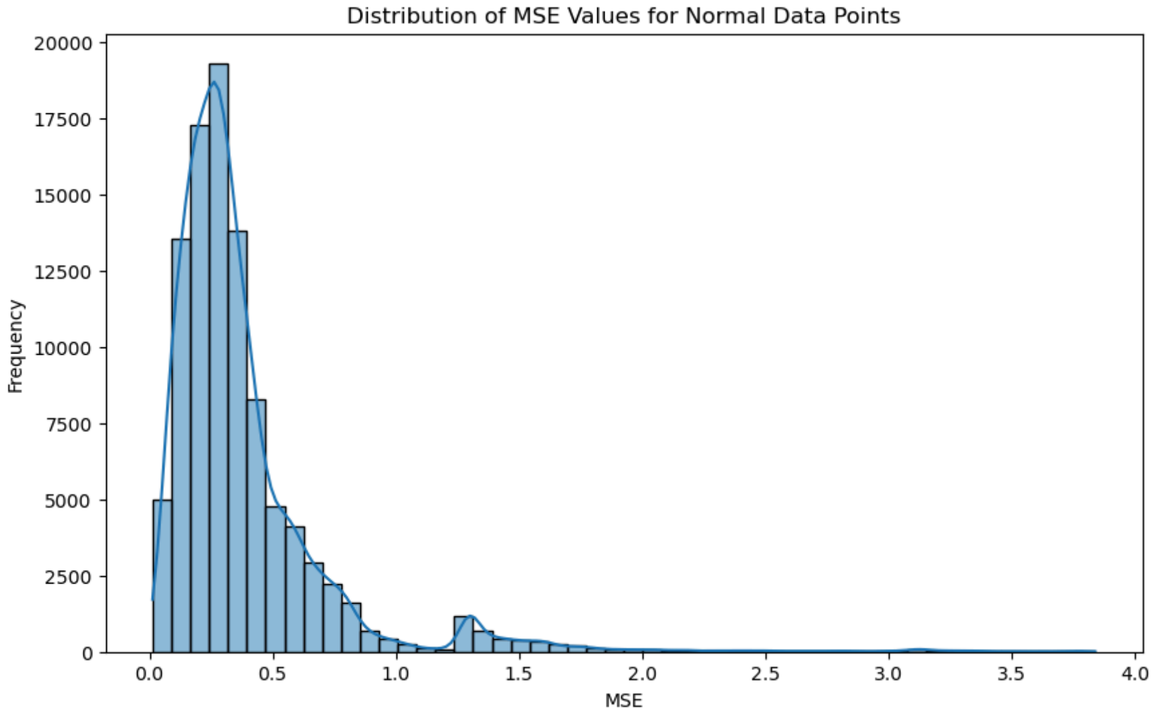


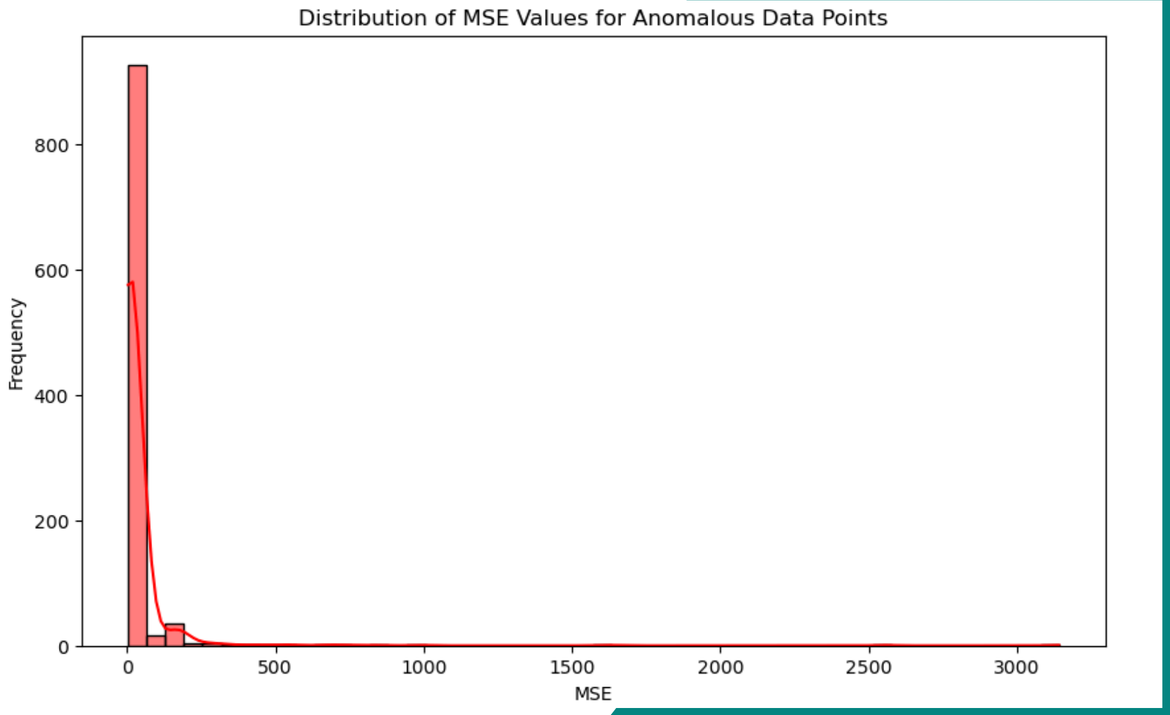


DL ALGORITHM

AUTO ENCODERS

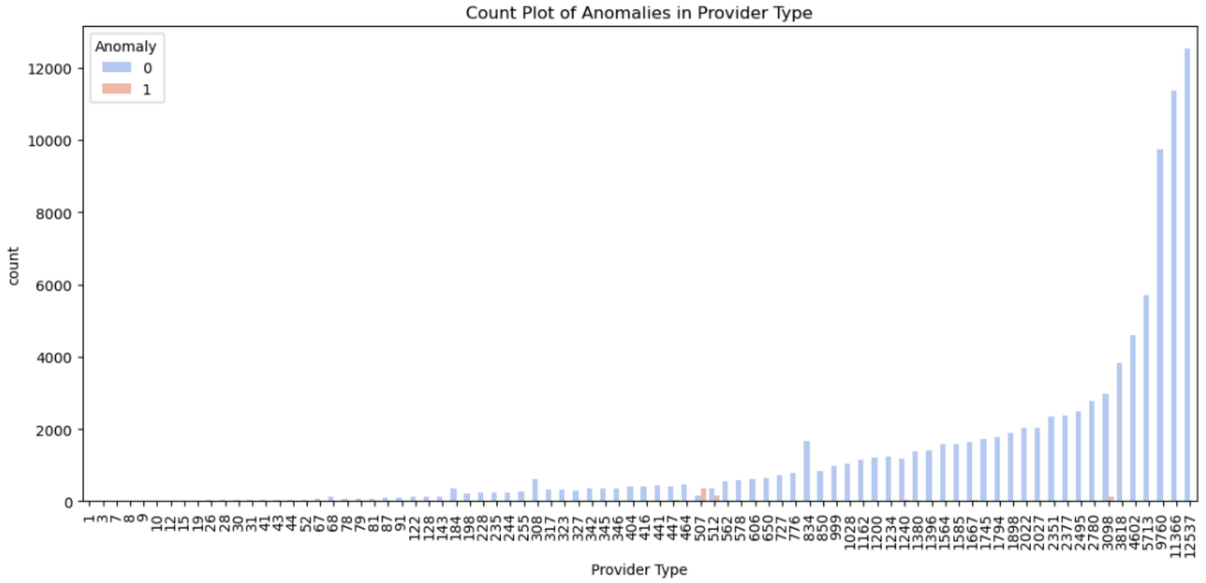
MSE VALUES

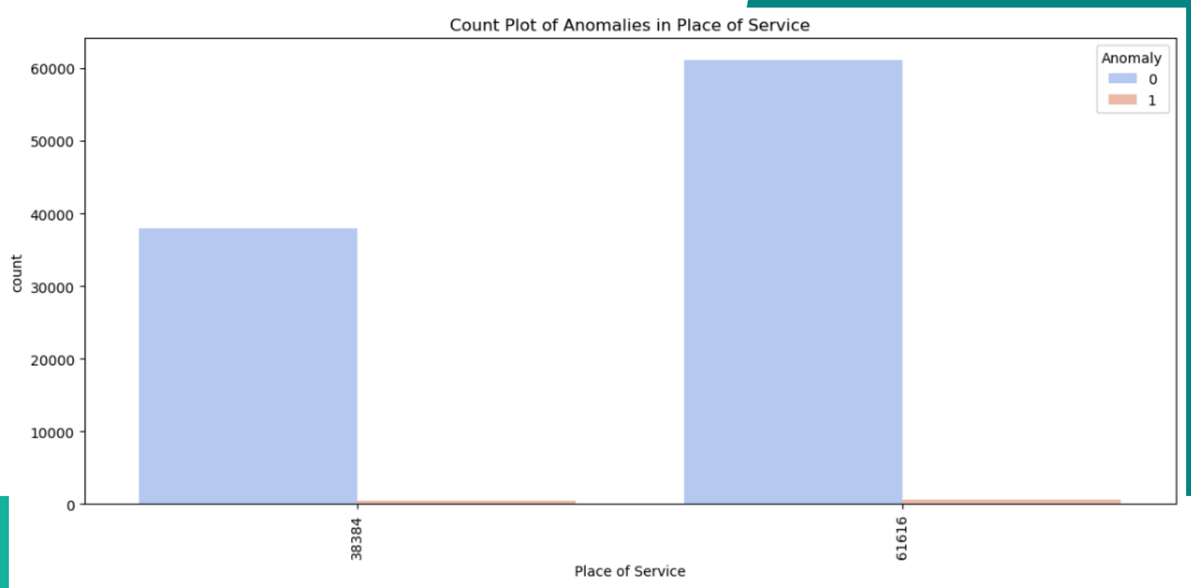




DL ALGORITHM

AUTO ENCODERS





CONCLUSION

In this project, we embarked on an exploration of healthcare provider data with the aim of detecting anomalies, particularly focusing on potential fraud detection.

Through extensive exploratory data analysis, we uncovered insightful patterns in provider credentials, geographic distributions, service metrics, and payment behaviors.

Utilizing machine learning algorithms including Isolation Forest and One-Class SVM, as well as deep learning techniques like autoencoders, we aimed to identify unusual patterns indicative of fraudulent activities.

The findings not only underscore the importance of robust data analytics in healthcare fraud detection but also highlight the potential of advanced machine learning methods in safeguarding healthcare integrity.

THANK YOU