Final Presentation: Infosys Springboard Internship 2024

Anomaly Detection in Healthcare Provider Data

Submitted by:
Rudrani Ghosh

Problem Statement

The goal is to detect anomalies in a healthcare provider dataset to identify potential fraudulent claims.

Dataset Details

The dataset contains 100,000 entries of healthcare provider's insurance claims data, including categorical and numerical features. There were initially 100,000 rows and 27 columns.

Some Categorical Columns: National Provider Identifier, Last Name, First Name, Middle Initial, Credentials, Gender, Entity Type, Street Address 1, Street Address 2, City, State Code, Postal Code of the Provider, HCPCS Code, HCPCS Description, HCPCS Drug Indicator

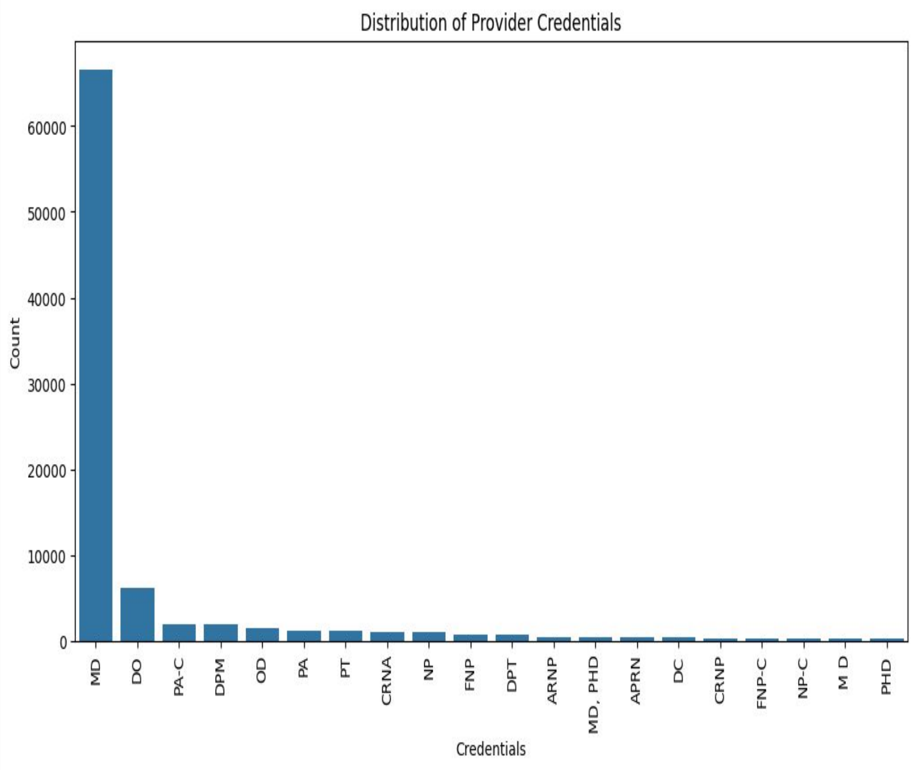
Numerical Columns: Number of Services, Number of Medicare Beneficiaries, Number of Distinct Medicare Beneficiary/Per Day Services, Average Medicare Allowed Amount, Average Submitted Charge Amount, Average Medicare Payment Amount, Average Medicare Standardized Amount

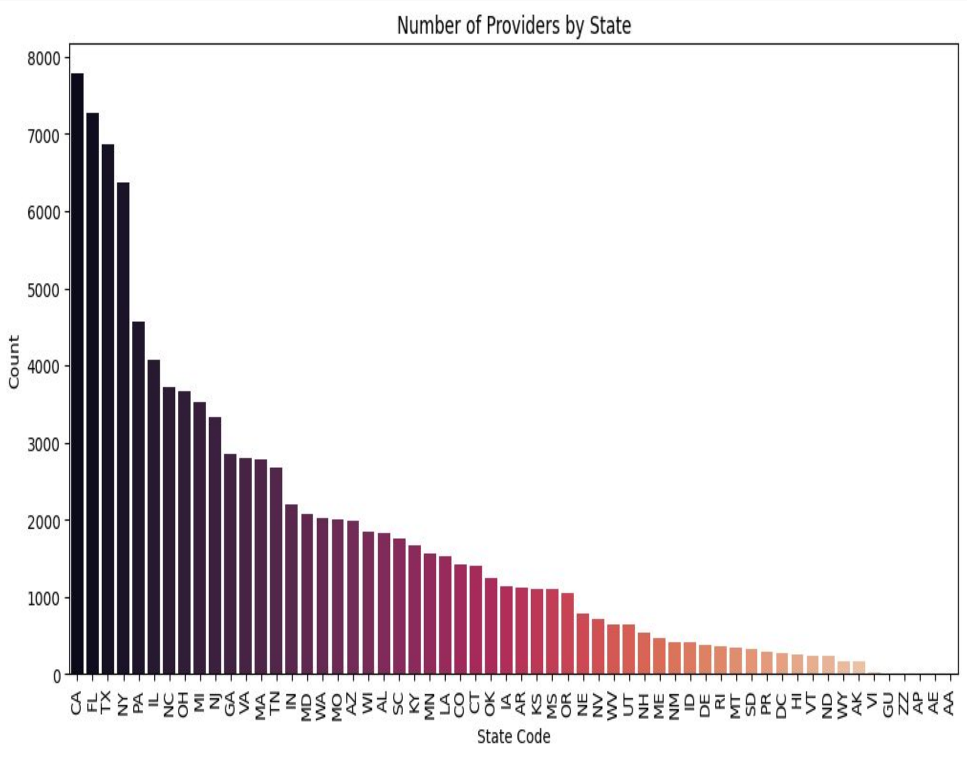
The final dataset after standardization consisted of 100000 rows and 11 columns

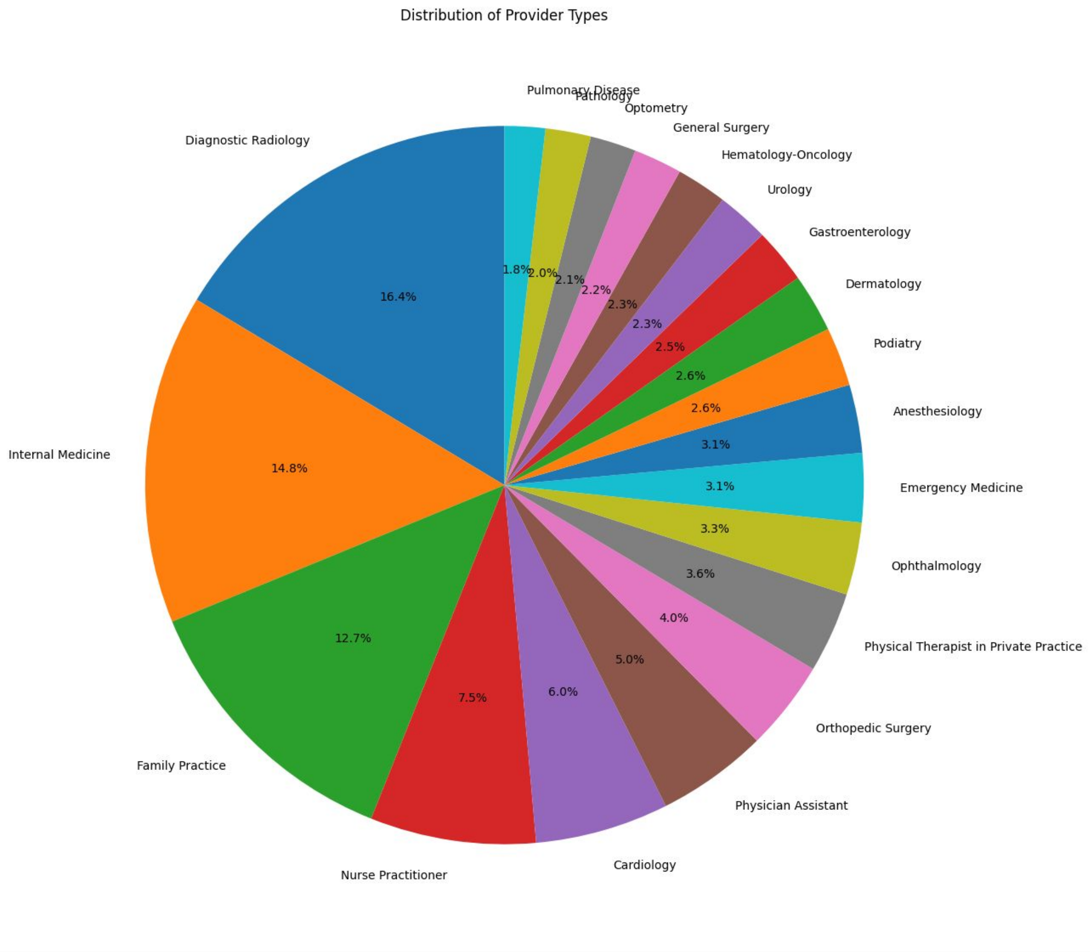
Preprocessing Steps

- Converting Object to Numeric Type
- Looking for Missing Values and imputing them with Mean
- Checking for Duplicate Values
- Merging the name columns into a single column- Full Name
- Merging the address columns
- Making the credentials column follow a standard nomenclature [MD is same as M.D. and so on]
- Frequency encoding categorical columns.
- Standardizing numerical columns.
- Dimensionality Reduction using PCA

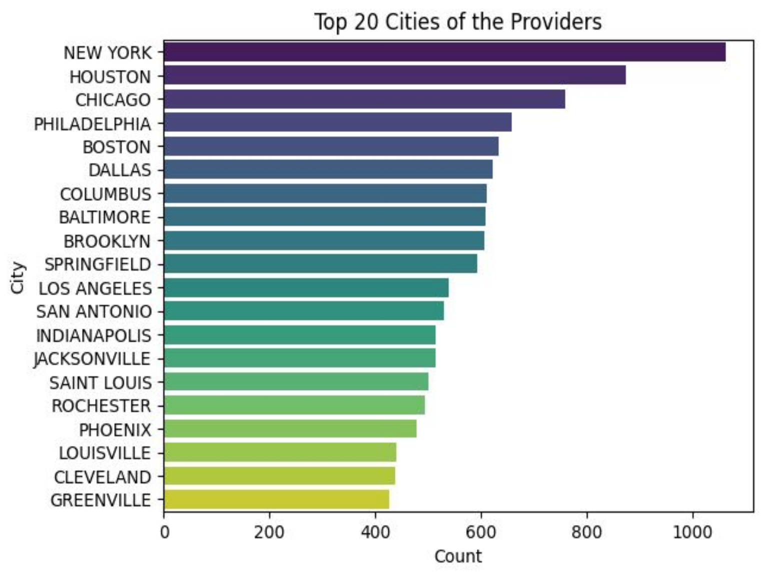
Exploratory Data Analysis Results:





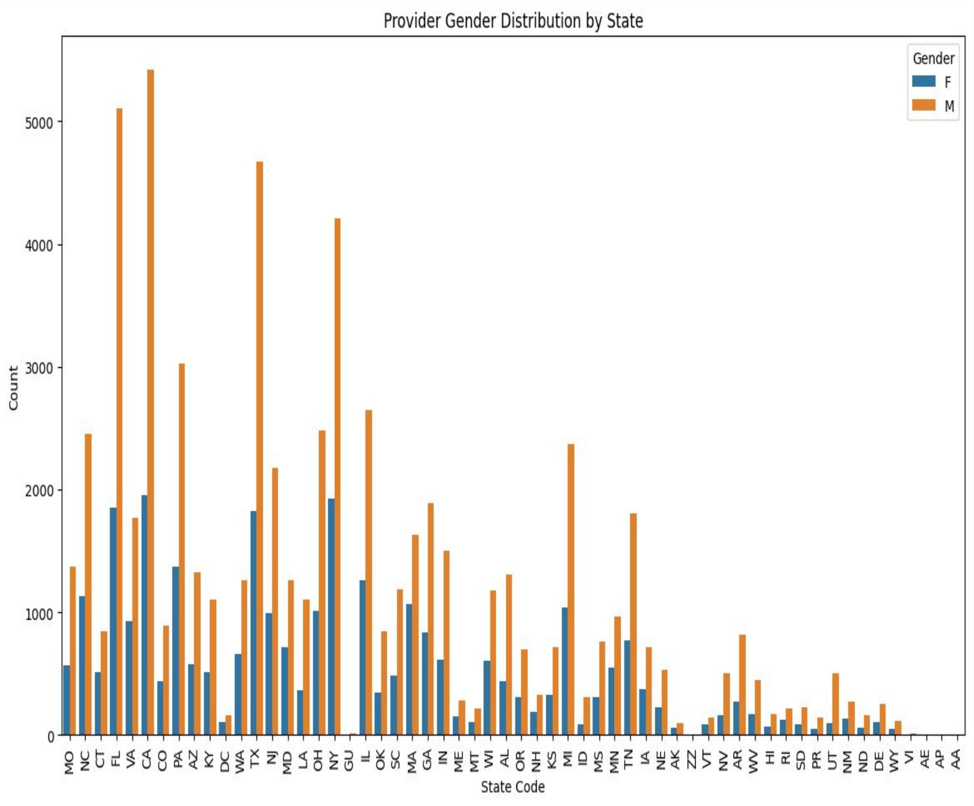


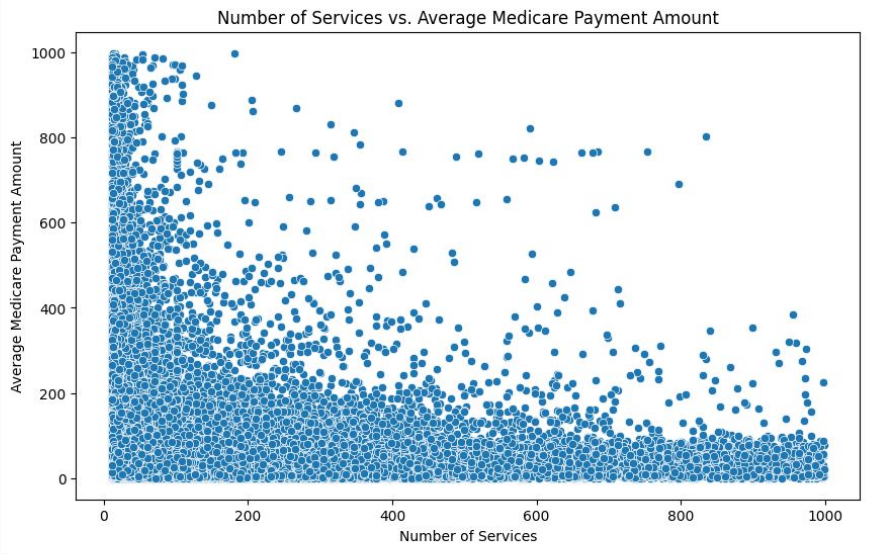
Distribution of Provider Types
Pulmonary Disease
Pathology
Optometry
General Surgery
Hematology-Oncology
Urology
Gastroenterology
Dermatology
Podiatry
Anesthesiology
Emergency Medicine
Ophthalmology
Physical Therapist in Private Practice
Orthopedic Surgery
Physician Assistant
Cardiology
Nurse Practitioner
Family Practice
Internal Medicine
Diagnostic Radiology
16.4%
14.8%
12.7%
7.5%
6.0%
5.0%
4.0%
3.6%
3.3%
3.1%
3.1%
2.6%
2.6%
2.5%
2.3%
2.3%
2.2%
2.1%
2.0%
1.8%

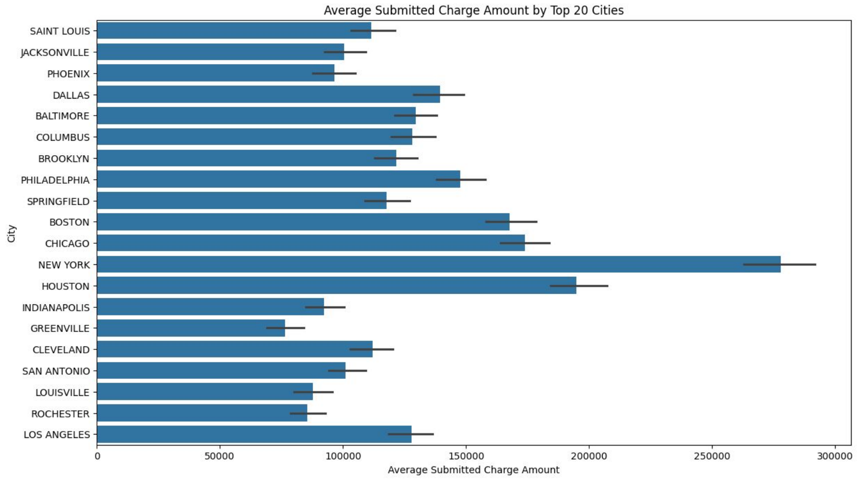


Top 20 Cities of the Providers
NEW YORK
HOUSTON
CHICAGO
PHILADELPHIA
BOSTON
DALLAS
COLUMBUS
BALTIMORE
BROOKLYN
SPRINGFIELD
LOS ANGELES
SAN ANTONIO
INDIANAPOLIS
JACKSONVILLE
SAINT LOUIS
ROCHESTER
PHOENIX
LOUISVILLE
CLEVELAND
GREENVILLE
City
0
200
400
600
800
1000
Count

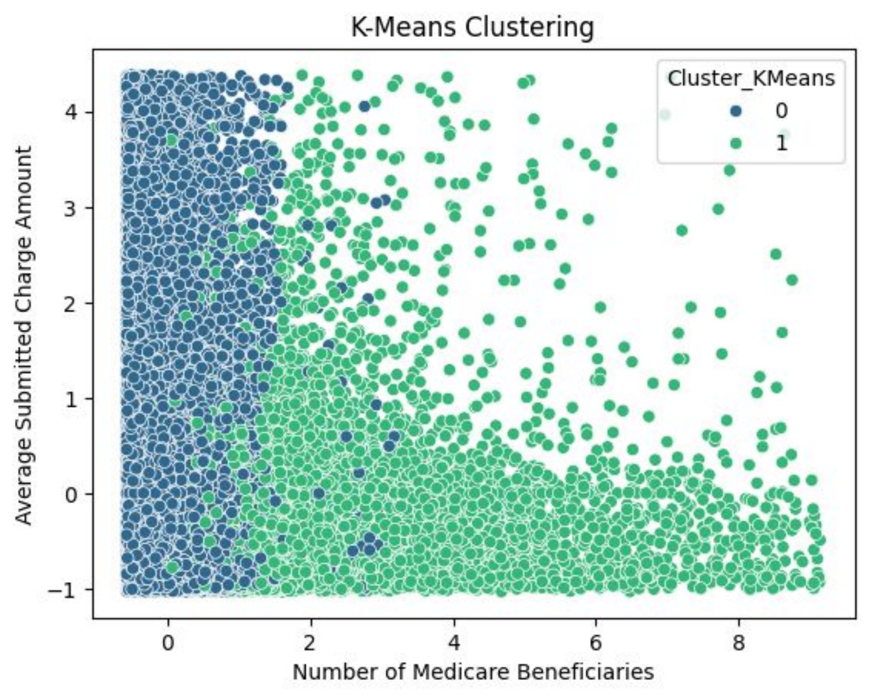
Bivariate Analysis

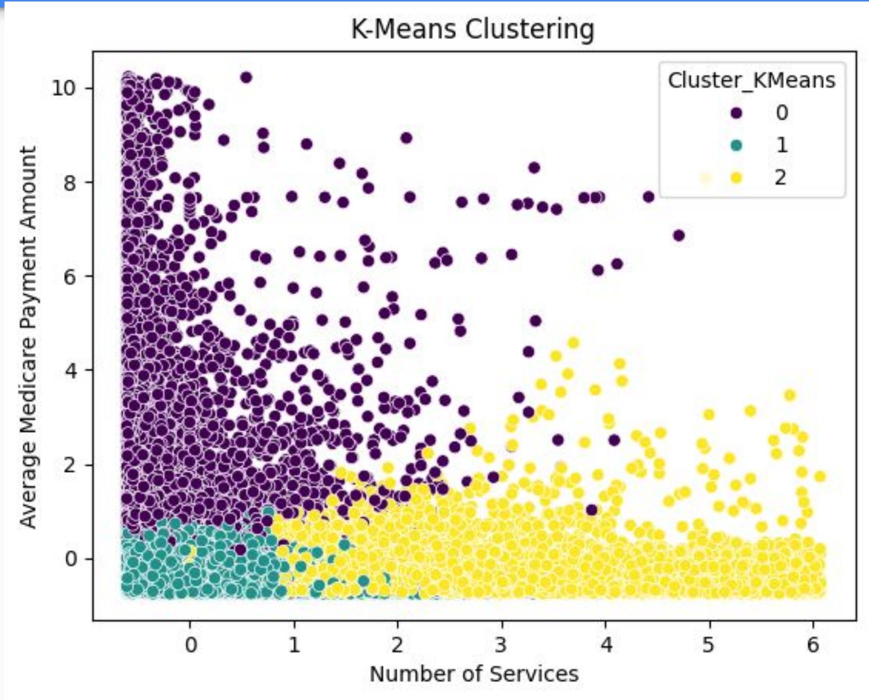




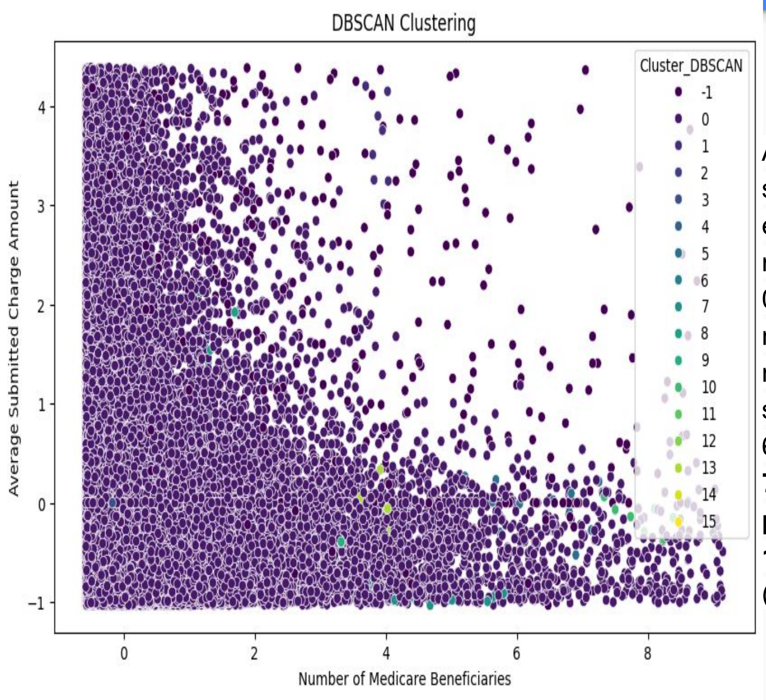


Clustering Results (K-Means)

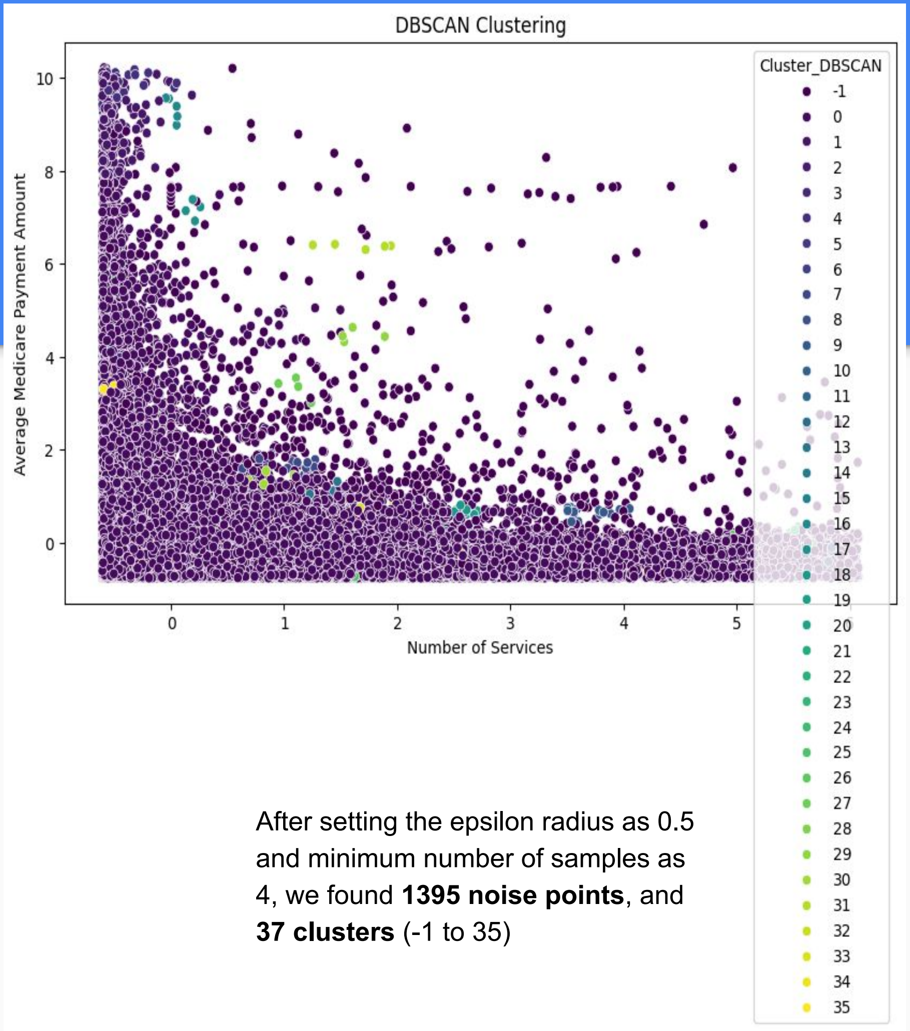




Clustering Results (DBScan)

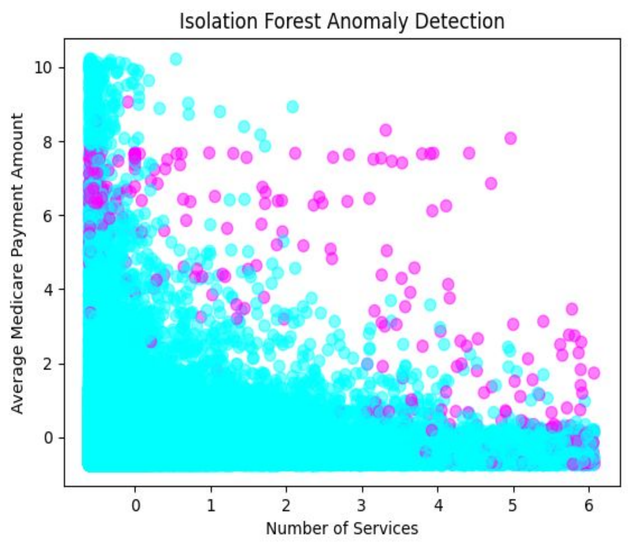


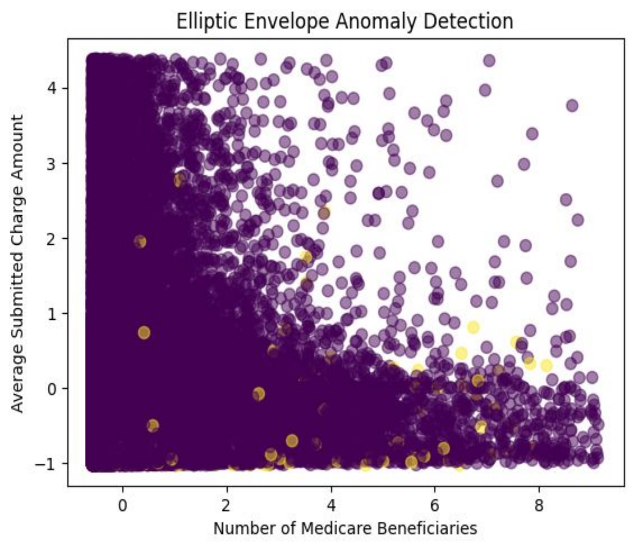
After setting the epsilon radius as 0.7 and minimum number of samples as 6, we found **788 noise points,** and **17 clusters** (-1 to 15)

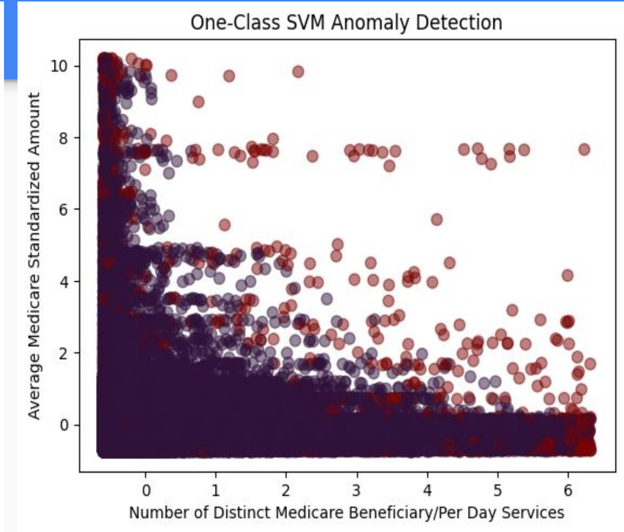


After setting the epsilon radius as 0.5 and minimum number of samples as 4, we found **1395 noise points**, and **37 clusters** (-1 to 35)

ML Algorithm Results





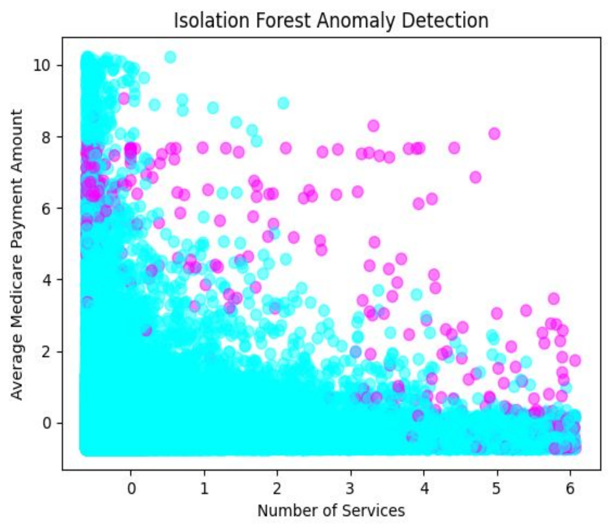


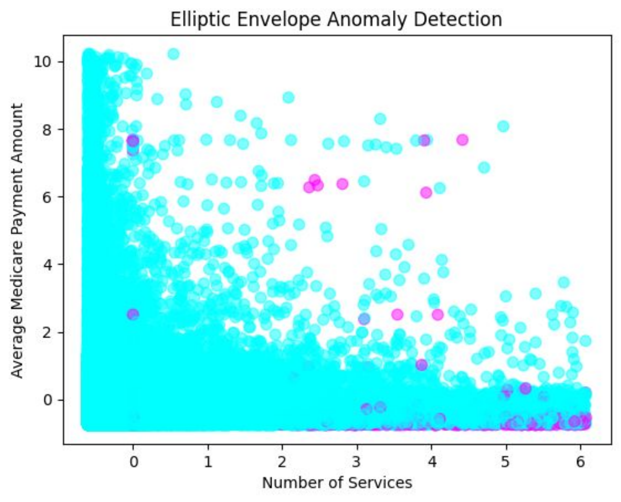
Using contamination as 0.005 and random_state as 0, the Isolation Forest model detected 500 anomalies

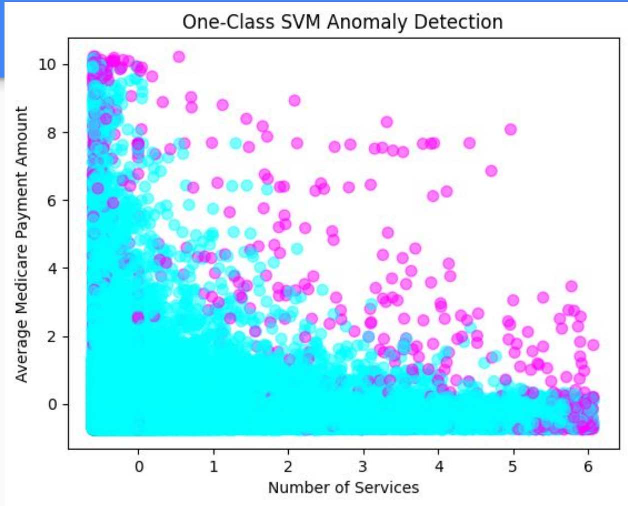
Using contamination as 0.007 and random_state as 42, the Elliptic Envelope model detected 700 anomalies

Using One-Class SVM model and setting gamma as 'auto' and nu as 0.01, 1012 anomalies have been detected

ML Algorithm Results using same columns





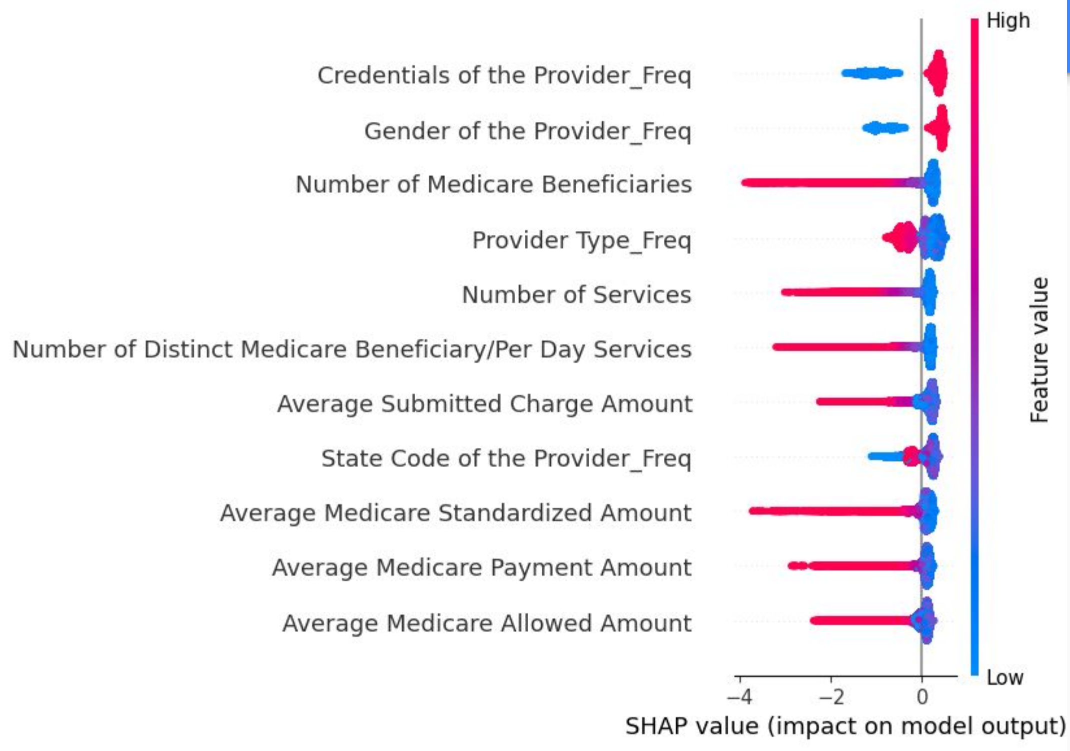


Using contamination as 0.005 and random_state as 0, the Isolation Forest model detected 500 anomalies

Using contamination as 0.007 and random_state as 42, the Elliptic Envelope model detected 700 anomalies

Using One-Class SVM model and setting gamma as 'auto' and nu as 0.01, 1012 anomalies have been detected

SHAP ANALYSIS OF Isolation Forest Model



INTERPRETATION

- The following columns tend to negatively affect the output:'Number of Services', 'Number of Medicare Beneficiaries', 'Number of Distinct Medicare Beneficiary/Per Day Services', 'Average Medicare Allowed Amount', 'Average Medicare Payment Amount', 'Average Medicare Standardized Amount',

- this shows the tendency of fraud increases with higher values in such columns

Deep Learning Results

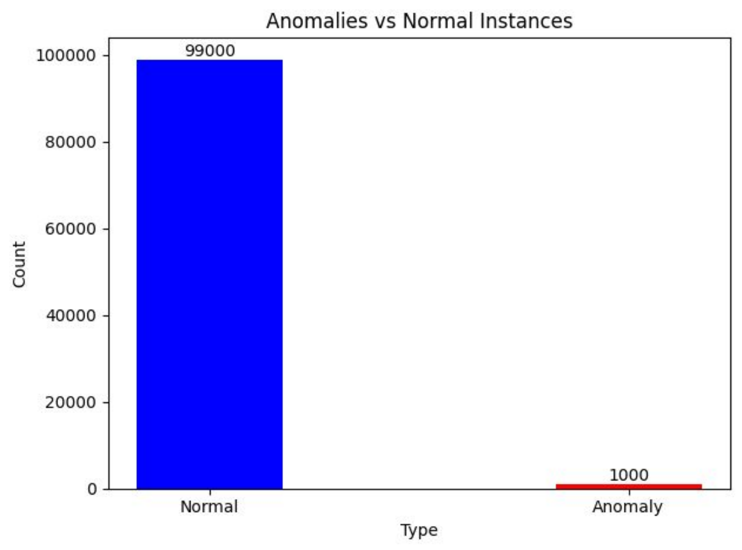
Model Training and Anomaly Detection Results

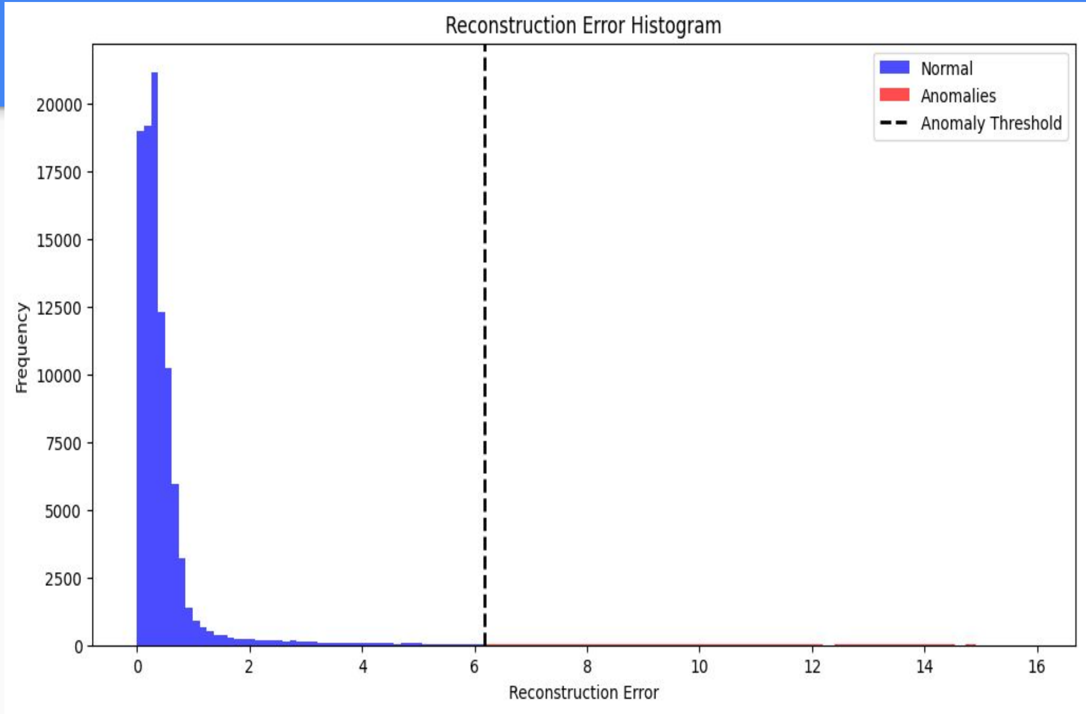
Training Progress:

- **Epochs Completed:** 3125

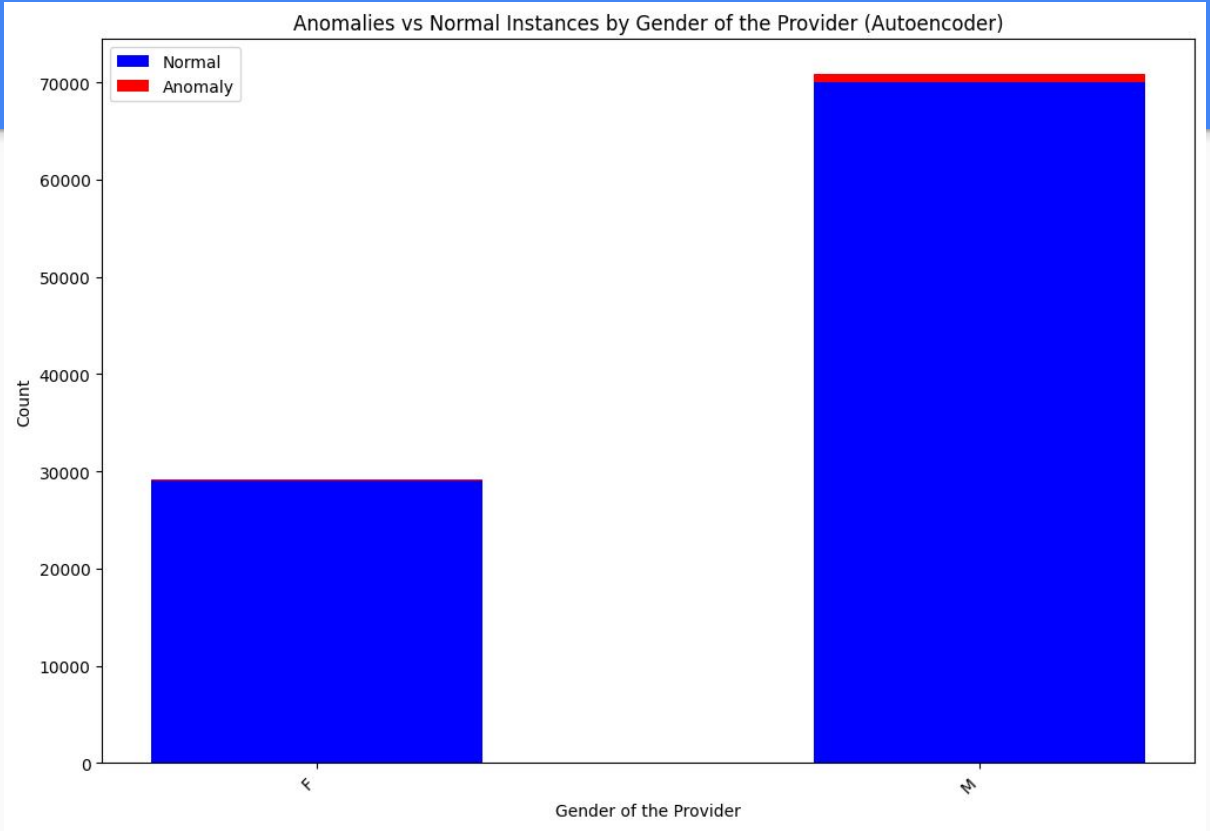
Anomaly Detection:

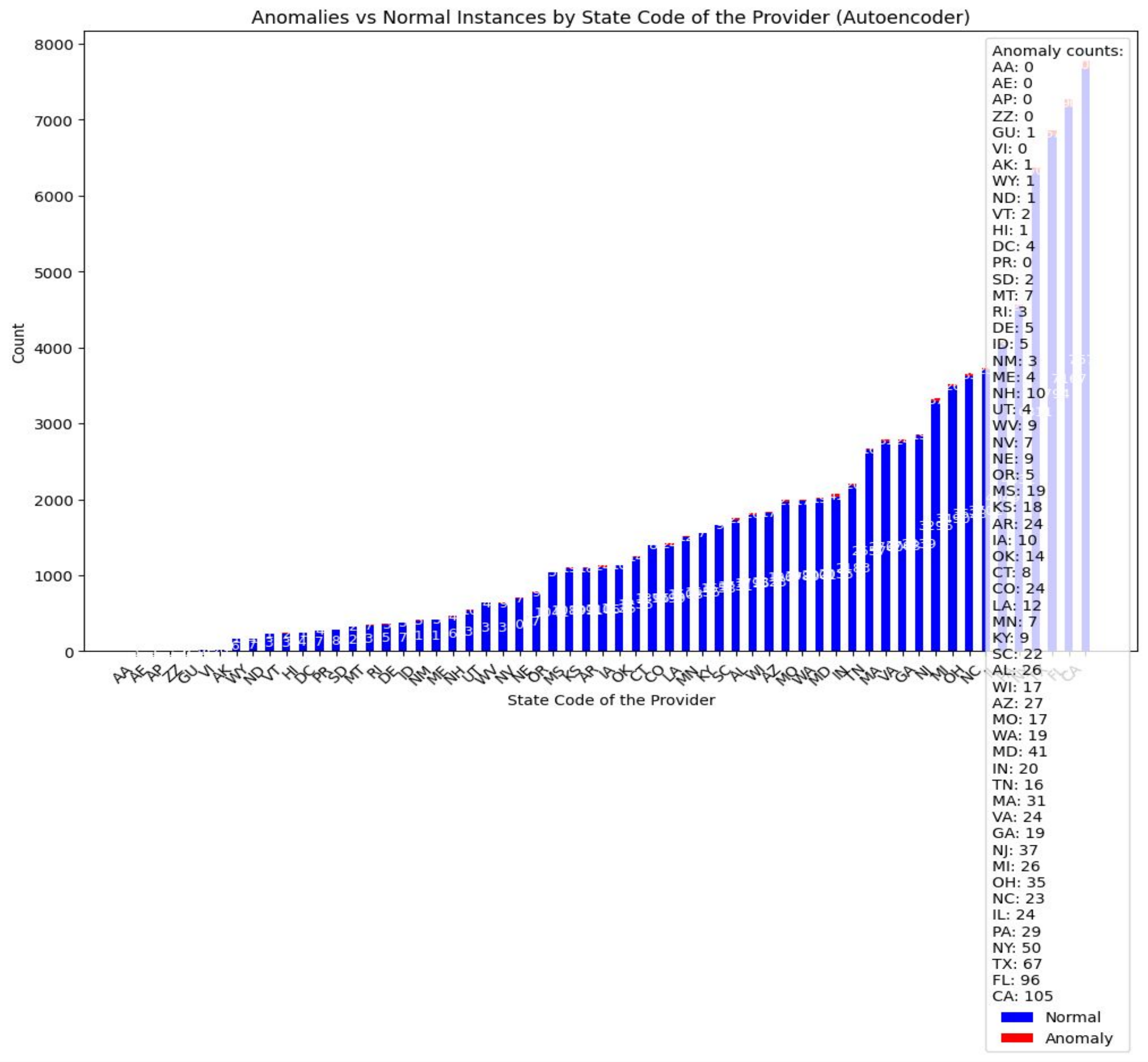
- **Number of Anomalies Detected:** 1000



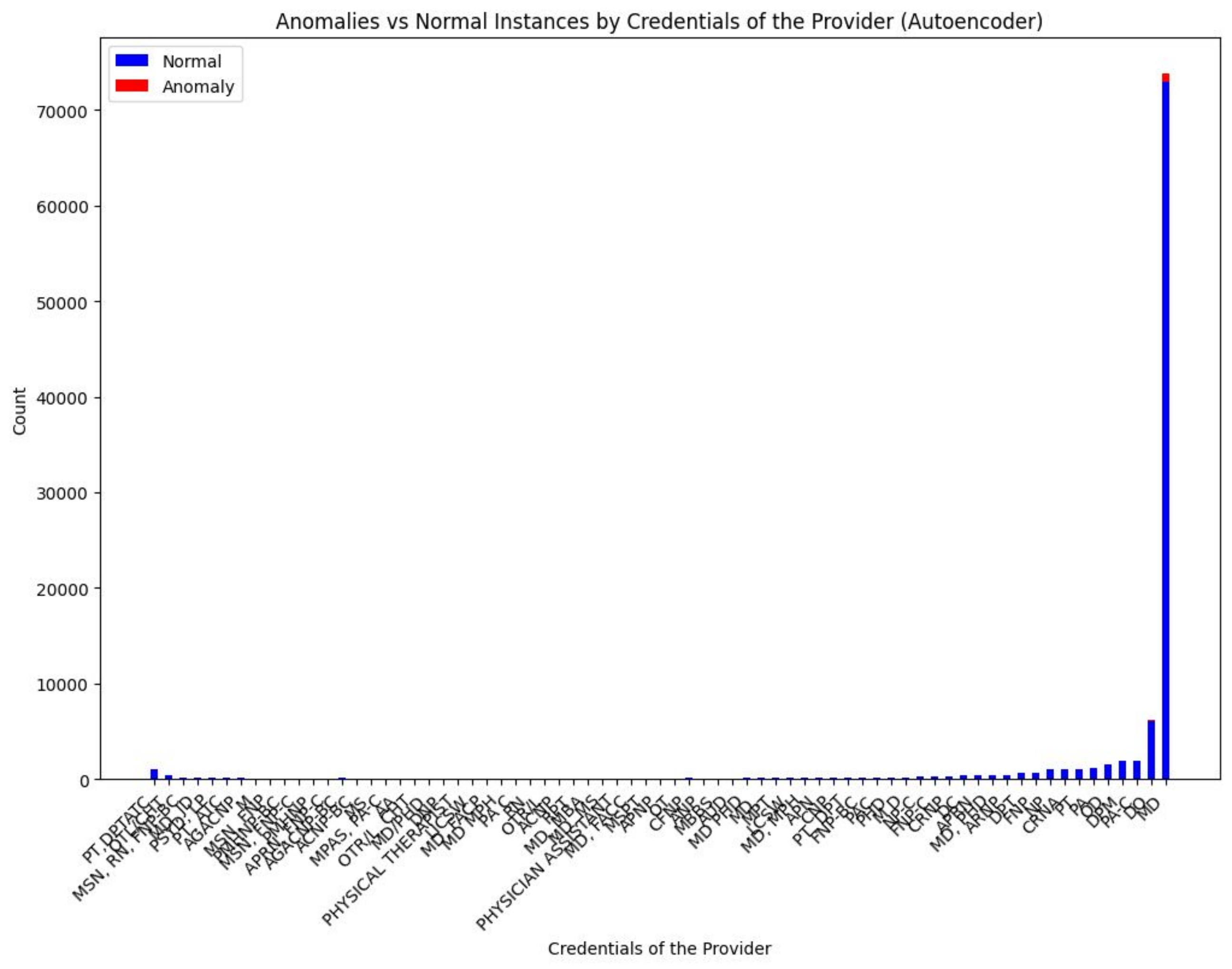


Deep Learning Results





Anomalies vs Normal Instances by State Code of the Provider (Autoencoder)
Count
State Code of the Provider
Anomaly counts:
AA: 0
AE: 0
AP: 0
ZZ: 0
GU: 1
VI: 0
AK: 1
WY: 1
ND: 1
VT: 2
HI: 1
DC: 4
PR: 0
SD: 2
MT: 7
RI: 3
DE: 5
ID: 5
NM: 3
ME: 4
NH: 10
UT: 4
WV: 9
NV: 7
NE: 9
OR: 5
MS: 19
KS: 18
AR: 24
IA: 10
OK: 14
CT: 8
CO: 24
LA: 12
MN: 7
KY: 9
SC: 22
AL: 26
WI: 17
AZ: 27
MO: 17
WA: 19
MD: 41
IN: 20
TN: 16
MA: 31
VA: 24
GA: 19
NJ: 37
MI: 26
OH: 35
NC: 23
IL: 24
PA: 29
NY: 50
TX: 67
FL: 96
CA: 105
Normal
Anomaly



Anomalies vs Normal Instances by Credentials of the Provider (Autoencoder)
Normal
Anomaly
Count
70000
60000
50000
40000
30000
20000
10000
0
Credentials of the Provider
MD

Model Architecture

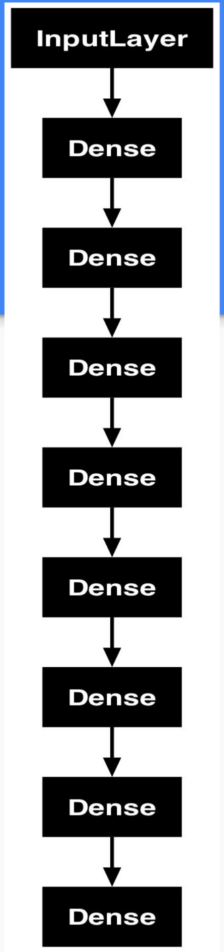
Model: "functional_1"

Layer (type)	Output Shape	Param #
input_layer_1 (InputLayer)	(None, 11)	0
dense_2 (Dense)	(None, 64)	768
dense_3 (Dense)	(None, 32)	2,080
dense_4 (Dense)	(None, 16)	528
dense_5 (Dense)	(None, 11)	187
dense_6 (Dense)	(None, 16)	192
dense_7 (Dense)	(None, 32)	544
dense_8 (Dense)	(None, 64)	2,112
dense_9 (Dense)	(None, 11)	715

Total params: 7,126 (27.84 KB)
Trainable params: 7,126 (27.84 KB)
Non-trainable params: 0 (0.00 B)

Interpretation:

The autoencoder consists of an encoding part (first four dense layers) and a decoding part (last four dense layers). The encoding layers reduce the input data to a lower-dimensional representation, while the decoding layers reconstruct the data back to its original dimensions. This structure helps the model learn an efficient representation of the input data, which can be useful for anomaly detection.



Comparison of Normal and Outlier Data MSE Values

```
Normal Data MSE: 0.463572730267268З
Anomaly Data MSE: 9.58481748853036
```

Normal Data MSE:

- **Value:** 0.464
- **Interpretation:** The MSE for normal data is 0.464. This indicates a low average reconstruction error for the data points that are not considered anomalies. The autoencoder performs well on the normal data, accurately reconstructing the input data with minimal error.

Anomaly Data MSE:

- **Value:** 9.585
- **Interpretation:** The MSE for anomaly data is 9.585. This significantly higher value compared to the normal data MSE suggests that the autoencoder struggles to reconstruct the anomalous data points accurately. The high reconstruction error confirms the presence of anomalies, highlighting that these data points differ substantially from the normal data.

Thank You