CALIFORNIA STATE UNIVERSITY
LONG BEACH

Drug Cost Prediction Analysis

IS 670 – Machine Learning for Business Analytics

Prof. Moy Patel

Group-2

Aditya Ravindra Chitnis (033718085)

Adhi Narayanan Ramesh (031536984)

Dhruti Dedhia (033694932)

Purav Rawal (032255195)

Vijay Krishna Ramraj (033696466)

Contents

1. Introduction……..1

2. Objectives……3

3. Dataset Intro……4

4. Data Understanding…….8

5. Data Cleaning…….12

6. Data Visualization ……13

7. Data preprocessing……23

8. Splitting the data……27

9. Model building & training……28

9.1 Random Forest Regressor……..28

9.2 K-Nearest Neighbors (KNN) Regressor……32

9.3 XGBoost Regressor……34

9.4 LightGBM Regressor…….35

9.5 Linear Regression Regressor……..38

9.6 Decision Tree Regressor…….40

10. Comparing models……..42

10. Conclusion & Discussion……..

Introduction

In today's data-driven healthcare landscape, the ability to **accurately predict drug costs** is critical not only for improving budget planning and identifying outliers, but also for supporting **policy-making** and ensuring the **financial sustainability** of public and private insurance systems. With prescription drug prices rising steadily, insurers and healthcare providers are under increasing pressure to analyse trends in drug usage and forecast future expenditures more effectively.

This project leverages real-world **insurance claims data** from the **Medicare Part D program**, which provides a detailed view of how much is spent on each drug annually. The dataset contains multi-year records (from 2018 to 2022) capturing **drug-level information** such as total spending, number of claims, dosage units dispensed, and number of beneficiaries. Analysing these metrics allows us to explore how drug costs change over time and how these patterns can inform both **forecasting and policy decisions** around reimbursement strategies and resource allocation. [1]

Using this dataset, we apply a wide range of **machine learning techniques** to develop models that predict the **average spending per dosage unit** — a key metric used by insurers and policymakers. Our analysis includes both supervised and unsupervised models such as **Linear Regression, Logistic Regression, K-Nearest Neighbors (KNN), K-Means Clustering, XGBoost, LightGBM**, and **Random Forest**. These models allow us to capture linear trends, detect non-linear patterns, and cluster similar drugs based on cost behaviours. [2]

In addition to model building, we also perform **correlation analysis, feature selection, scaling, and hyperparameter tuning** to improve performance and interpretability. The ultimate goal of this study is to evaluate the effectiveness of these algorithms in forecasting drug unit costs, compare their predictive accuracy, and provide **actionable insights** that can guide **insurance reimbursement strategies**, improve **cost monitoring**, and support **data-driven policy decisions** in the pharmaceutical sector. [2]

Objectives

- **Understand Drug Spending Patterns from a Vast Multi-Year Dataset**
  To analyse a large-scale dataset of Medicare Part D insurance claims covering five years (2018–2022), and explore trends in drug-level spending, claims, and beneficiary behavior across this extensive time span.
- **Identify Key Cost Drivers in Drug Claims**
  To determine which historical features (e.g., spending per claim, dosage unit, or beneficiary) most strongly influence future drug costs using correlation analysis and feature selection.
- **Predict Average Drug Spending Per Dosage Unit**
  To develop machine learning models that accurately forecast Avg_Spnd_Per_Dsg_Unt_Wghtd_2022, enabling proactive cost management and decision-making for healthcare payers.
- **Support Policy and Reimbursement Planning**
  To provide data-driven insights that can inform prescription benefit design, resource allocation, and pricing negotiations in insurance and policy frameworks.
- **Apply and Compare Multiple Machine Learning Algorithms**
  To train, evaluate, and compare the performance of various models — including Linear Regression, Logistic Regression, KNN, XGBoost, Random Forest, LightGBM, and K-Means — in predicting drug costs or classifying outliers.
- **Detect and Address Outliers in Drug Spending**
  To identify cost outliers within drug claims and apply appropriate preprocessing techniques such as log transformation to improve model robustness.
- **Visualize Insights for Interpretation and Reporting**
  To create clear and informative visualizations that highlight important patterns, support model interpretations, and communicate findings to stakeholders.

Dataset Intro

This project uses a comprehensive, multi-year dataset that captures key spending and usage patterns related to prescription drugs under a public healthcare system. The dataset provides detailed information about individual drugs, including total spending, dosage units, claims, and average spending per dosage unit — tracked annually from **2018 to 2022**. [1]

This rich dataset enables the identification of trends in drug costs over time, which is essential for building predictive models. With over **13,800 records** and **46 attributes**, the data is well-suited for both exploratory analysis and machine learning applications. It offers a solid foundation for understanding cost behavior and constructing models that can estimate future drug pricing.[1] The details of the dataset are:

Dataset Title: Medicare Part D Spending by Drug

Number of Entries: 13,889

Number of Attributes: 46

Table 1: description of dataset

No.	Attribute Name	Description
1	Brnd_Name	Brand name of the drug
2	Gnrc_Name	Generic name of the drug
3	Tot_Mftr	Total number of manufacturers for the drug
4	Mftr_Name	Name of the manufacturer
5	Tot_Spndng_2018	Total spending in

		2018
6	Tot_Dsg_Unts_2018	Total dosage units dispensed in 2018
7	Tot_Clms_2018	Total number of claims in 2018
8	Tot_Benes_2018	Total number of beneficiaries in 2018
9	Avg_Spnd_Per_Dsg_Unt_Wghtd_2018	Average spending per dosage unit in 2018
10	Avg_Spnd_Per_Clm_2018	Average spending per claim in 2018
11	Avg_Spnd_Per_Bene_2018	Average spending per beneficiary in 2018
12	Outlier_Flag_2018	Outlier flag for 2018
13	Tot_Spndng_2019	Total spending in 2019
14	Tot_Dsg_Unts_2019	Total dosage units dispensed in 2019
15	Tot_Clms_2019	Total number of claims2019
16	Tot_Benes_2019	Total number of beneficiaries in 2019
17	Avg_Spnd_Per_Dsg_Unt_Wghtd_2019	Average spending per dosage unit

		(weighted) in 2019
18	Avg_Spnd_Per_Clm_2019	Average spending per claim in 2019
19	Avg_Spnd_Per_Bene_2019	Average spending per beneficiary in 2019
20	Outlier_Flag_2019	Outlier flag for 2019
21	Tot_Spndng_2020	Total spending in 2020
22	Tot_Dsg_Unts_2020	Total dosage units dispensed in 2020
23	Tot_Clms_2020	Total number of claims in 2020
24	Tot_Benes_2020	Total number of beneficiaries in 2020
25	Avg_Spnd_Per_Dsg_Unt_Wghtd_2020	Average spending per dosage unit (weighted) in 2020
26	Avg_Spnd_Per_Clm_2020	Average spending per claim in 2020
27	Avg_Spnd_Per_Bene_2020	Average spending per beneficiary in 2020
28	Outlier_Flag_2020	Outlier flag for 2020
29	Tot_Spndng_2021	Total spending in 2021

30	Tot_Dsg_Unts_2021	Total dosage units dispensed 31in 2021
31	Tot_Clms_2021	Total number of claims in 2021
32	Tot_Benes_2021	Total number of beneficiaries in 2021
33	Avg_Spnd_Per_Dsg_Unt_Wghtd_2021	Average spending per dosage unit (weighted) in 2021
34	Avg_Spnd_Per_Clm_2021	Average spending per claim in 2021
35	Avg_Spnd_Per_Bene_2021	Average spending per beneficiary in 2021
36	Outlier_Flag_2021	Outlier flag for 2021
37	Tot_Spndng_2022	Total spending in 2022
38	Tot_Dsg_Unts_2022	Total dosage units dispensed in 2022
39	Tot_Clms_2022	Total number of claims in 2022
40	Tot_Benes_2022	Total number of beneficiaries in 2022
41	Avg_Spnd_Per_Dsg_Unt_Wghtd_2022	Target variable-average spending per dosage unit

		(weighted) in 2022
42	Avg_Spnd_Per_Clm_2022	Average spending per claim in 2022
43	Avg_Spnd_Per_Bene_2022	Average spending per beneficiary in 2022
44	Outlier_Flag_2022	Outlier flag for 2022
45	Chg_Avg_Spnd_Per_Dsg_Unt_21_22	Change in average spending per dosage unit from 2021 to 2022
46	CAGR_Avg_Spnd_Per_Dsg_Unt_18_22	Compound annual growth rate of spending per dosage unit (2018 - 2022)

Data Understanding

To begin the project, all essential Python libraries such as Pandas, NumPy, Seaborn, Matplotlib, and Scikit-learn were imported into a Jupyter Notebook environment. These libraries provided the tools necessary for data handling, visualization, preprocessing, and machine learning model development.

```
# Basic data handling
import pandas as pd
import numpy as np

# Visualization libraries
import matplotlib.pyplot as plt
import seaborn as sns
import missingno as msno

# Machine learning preprocessing and evaluation
from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn.metrics import r2_score, mean_absolute_error, mean_squared_error
from sklearn.neighbors import KNeighborsRegressor
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestRegressor
from sklearn.cluster import KMeans

# Advanced models
import xgboost as xgb
import lightgbm as lgb
```

Figure 1: Importing all the necessary libraries

The dataset was loaded into a DataFrame using the read_csv() function from the pandas library. This function reads the contents of the CSV file and returns a structured two-dimensional table, allowing for easy access and analysis of the data. A preview of the first few rows is shown below:

	Brnd_Name	Gnrc_Name	Tot_Mftr	Mftr_Name	Tot_Spndng_2018	Tot_Dsg_Unts_2018	Tot_Clms_2018	Tot_Benes_2018	Avg_Spnd_Per_Dsg_Unt_Wghtd_2018	Avg_Spnd_Per_
0	1st Tier Unifine Pentips	Pen Needle, Diabetic	1	Overall	167193.78	761658.0	6538.0	2341.0	0.219785	2
1	1st Tier Unifine Pentips	Pen Needle, Diabetic	1	Owen Mumford Us	167193.78	761658.0	6538.0	2341.0	0.219785	2
2	1st Tier Unifine Pentips Plus	Pen Needle, Diabetic	1	Overall	369402.85	1813908.0	14931.0	5674.0	0.203658	2
3	1st Tier Unifine Pentips Plus	Pen Needle, Diabetic	1	Owen Mumford Us	369402.85	1813908.0	14931.0	5674.0	0.203658	2
4	Abacavir	Abacavir Sulfate	6	Overall	10653423.32	3034767.0	40388.0	7359.0	4.032155	26

rows × 46 columns

Figure 2: Reading the dataset

```
df.shape

(13889, 46)
```

Figure 3: Shape of the dataset

No.	Attribute Name	Description	Data Type
1	Brnd_Name	Brand name of the drug	object
2	Gnrc_Name	Generic name of the drug	object
3	Tot_Mftr	Total number of manufacturers for the drug	int64
4	Mftr_Name	Manufacturer name	object
5	Tot_Spndng_2018	Total spending in 2018	float64
6	Tot_Dsg_Unts_2018	Total dosage units dispensed in 2018	float64
7	Tot_Clms_2018	Total number of claims in 2018	float64
8	Tot_Benes_2018	Total number of beneficiaries in 2018	float64
9	Avg_Spnd_Per_Dsg_Unt_Wghtd_2018	Average spending per dosage unit in 2018	float64
10	Avg_Spnd_Per_Clm_2018	Average spending per claim in 2018	float64
11	Avg_Spnd_Per_Bene_2018	Average spending per beneficiary in 2018	float64
12	Outlier_Flag_2018	Outlier flag for 2018	float64
13	Tot_Spndng_2019	Total spending in 2019	float64
14	Tot_Dsg_Unts_2019	Total dosage units dispensed in	float64

		2019	
15	Tot_Clms_2019	Total number of claims in 2019	float64
16	Tot_Benes_2019	Total number of beneficiaries in 2019	float64
17	Avg_Spnd_Per_Dsg_Unt_Wghtd_2019	Average spending per dosage unit in 2019	float64
18	Avg_Spnd_Per_Clm_2019	Average spending per claim in 2019	float64
19	Avg_Spnd_Per_Bene_2019	Average spending per beneficiary in 2019	float64
20	Outlier_Flag_2019	Outlier flag for 2019	float64
21	Tot_Spndng_2020	Total spending in 2020	float64
22	Tot_Dsg_Unts_2020	Total dosage units dispensed in 2020	float64
23	Tot_Clms_2020	Total number of claims in 2020	float64
24	Tot_Benes_2020	Total number of beneficiaries in 2020	float64
25	Avg_Spnd_Per_Dsg_Unt_Wghtd_2020	Average spending per dosage unit in 2020	float64
26	Avg_Spnd_Per_Clm_2020	Average spending per claim in 2020	float64
27	Avg_Spnd_Per_Bene_2020	Average spending per beneficiary in 2020	float64
28	Outlier_Flag_2020	Outlier flag for 2020	float64
29	Tot_Spndng_2021	Total spending in 2021	float64
30	Tot_Dsg_Unts_2021	Total dosage units dispensed in 2021	float64
31	Tot_Clms_2021	Total number of claims in 2021	float64
32	Tot_Benes_2021	Total number of beneficiaries in 2021	float64

33	Avg_Spnd_Per_Dsg_Unt_Wghtd_2021	Average spending per dosage unit in 2021	float64
34	Avg_Spnd_Per_Clm_2021	Average spending per claim in 2021	float64
35	Avg_Spnd_Per_Bene_2021	Average spending per beneficiary in 2021	float64
36	Outlier_Flag_2021	Outlier flag for 2021	float64
37	Tot_Spndng_2022	Total spending in 2022	float64
38	Tot_Dsg_Unts_2022	Total dosage units dispensed in 2022	float64
39	Tot_Clms_2022	Total number of claims in 2022	int64
40	Tot_Benes_2022	Total number of beneficiaries in 2022	float64
41	Avg_Spnd_Per_Dsg_Unt_Wghtd_2022	Target variable - average spending per dosage unit in 2022	float64
42	Avg_Spnd_Per_Clm_2022	Average spending per claim in 2022	float64
43	Avg_Spnd_Per_Bene_2022	Average spending per beneficiary in 2022	float64
44	Outlier_Flag_2022	Outlier flag for 2022	float64
45	Chg_Avg_Spnd_Per_Dsg_Unt_21_22	Change in average spending per dosage unit from 2021 to 2022	float64
46	CAGR_Avg_Spnd_Per_Dsg_Unt_18_22	Compound annual growth rate of spending per dosage unit (2018-2022)	float64

Table 2: Dataset Information

After successfully loading the dataset into a DataFrame, we reviewed its structure and contents. The shape of the dataset is **(13,889, 46)**, indicating it contains **13,889 rows**

(samples) and **46 columns (features)**. This allowed us to understand how the drug data is organized across multiple years, covering both raw and derived metrics.

Upon closer examination, the dataset contains two main types of features:

Numerical Features:

Most of the dataset consists of numerical values that represent various spending and usage metrics. Some of the key numerical features include:

1. Tot_Spndng_2022
2. Tot_Dsg_Unts_2022
3. Tot_Clms_2022
4. Avg_Spnd_Per_Dsg_Unt_Wghtd_2022
5. Avg_Spnd_Per_Clm_2021
6. Avg_Spnd_Per_Bene_2020
7. CAGR_Avg_Spnd_Per_Dsg_Unt_18_22
8. All similar columns from 2018–2021

Categorical Features:

Only a few columns are non-numeric, mainly descriptive labels:

1. Brnd_Name
2. Gnrc_Name
3. Mftr_Name

Data Cleaning

To prepare the dataset for modelling, we performed several cleaning steps to handle missing and duplicate values:

1. **Dropped columns with more than 50% missing data** — These columns lacked sufficient information and were removed to maintain dataset quality.
2. **Removed rows with any remaining null values** — Ensuring the dataset was complete and ready for training without requiring imputation.
3. **Eliminated duplicate entries** — Prevented duplicate records from skewing results or introducing bias.

These steps helped reduce noise and ensured that the dataset used in the analysis was reliable and consistent.

```
# ---------------------- DATA CLEANING ----------------------

# Drop columns with more than 50% missing values
df_cleaned = df.loc[:, df.isnull().mean() < 0.5]

# Drop rows with any missing values
df_cleaned = df_cleaned.dropna()

# Drop duplicate rows
df_cleaned = df_cleaned.drop_duplicates()
```

Figure 5: Data cleaning steps applied to remove nulls, incomplete rows, and duplicates

Data visualization

We expanded our analysis beyond a single year by evaluating the proportion of drugs flagged as outliers across all five years from 2018 to 2022. The chart illustrates the year-wise distribution of outlier (Outlier_Flag = 1) and non-outlier (Outlier_Flag = 0) drugs based on cost behavior.

- In each year, over 90% of drugs were classified as non-outliers, indicating stable pricing for the majority.
- A consistent 6–8% of drugs were flagged as outliers, suggesting the presence of high-cost drugs every year.
- The outlier trend remained relatively stable over time, which supports the reliability of using historical patterns for future cost prediction.

This multi-year perspective is important for understanding **cost volatility** and ensuring robust **outlier detection strategies** when forecasting drug prices.

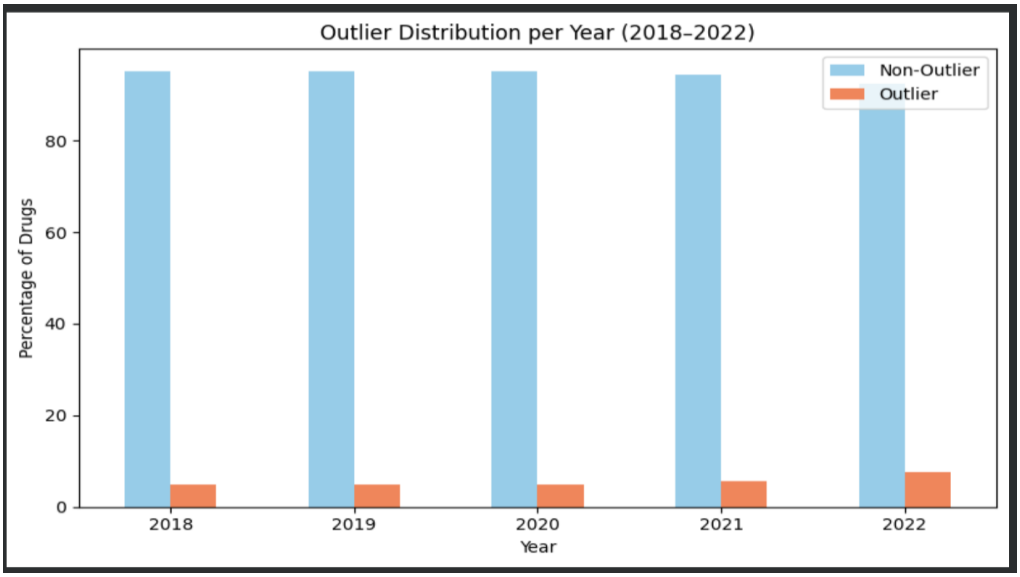


Figure 6: Year-wise bar chart showing the distribution of outlier vs. non-outlier drugs (2018–2022)

This grouped bar chart illustrates the distribution of the **top five most frequently occurring drug brands** over a five-year period (2018–2022). The count is based on the number of non-null claim records (Tot_Clms_YYYY) for each year, representing how often each drug appeared in Medicare Part D data.

- Insulin Syringe consistently leads across all five years, indicating its continued high usage and prescribing frequency.
- Brands such as Fenofibrate, Levetiracetam, Potassium Chloride, and Gabapentin also maintain steady presence, though with varying trends year-to-year.
- This multi-year perspective highlights long-term prescribing patterns for high-volume drugs and helps identify persistent cost drivers in the pharmaceutical space.

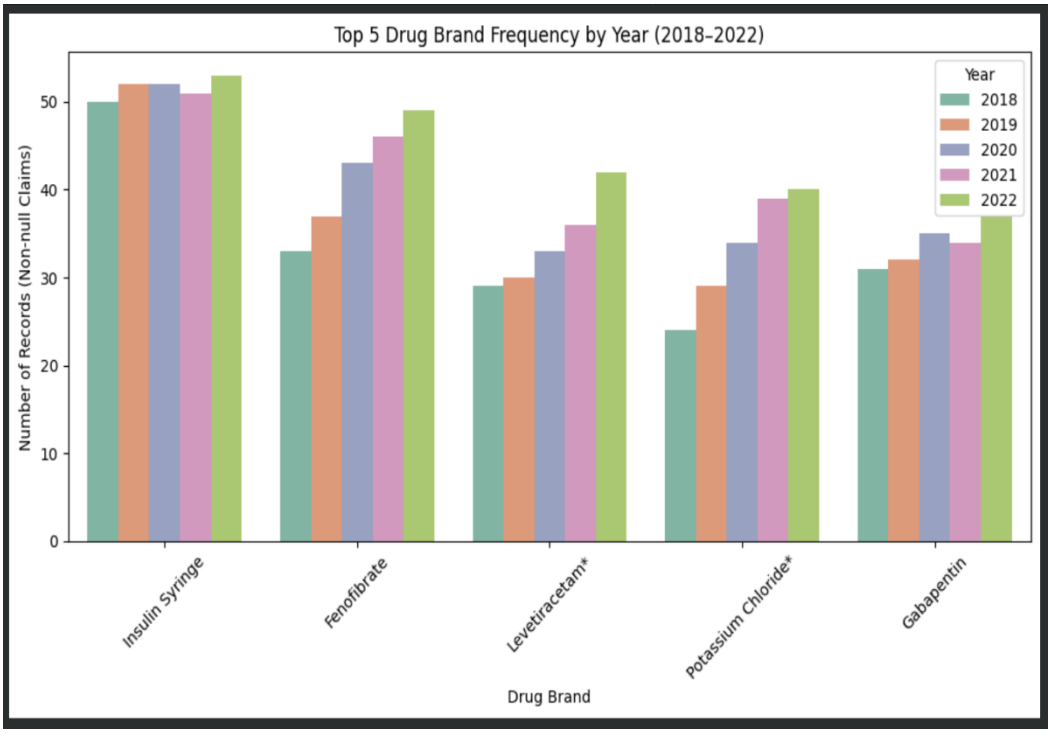


Figure 7: Grouped bar chart showing yearly frequency of the top 5 drug brands (2018–2022)

This grouped bar chart visualizes the **total outlier vs. non-outlier classifications** across all five years (2018–2022) for the top five most frequently occurring drug brands.

- Insulin Syringe and Levetiracetam show higher instances of outliers compared to other drugs.
- Gabapentin and Potassium Chloride remained consistently classified as non-outliers across all years.
- The chart highlights which drugs are most prone to cost anomalies over time and reinforces the stability of others.

This multi-year analysis provides a **long-term view of cost behavior**, helping identify **reliable vs. volatile drugs** that may influence healthcare pricing and budgeting decisions.

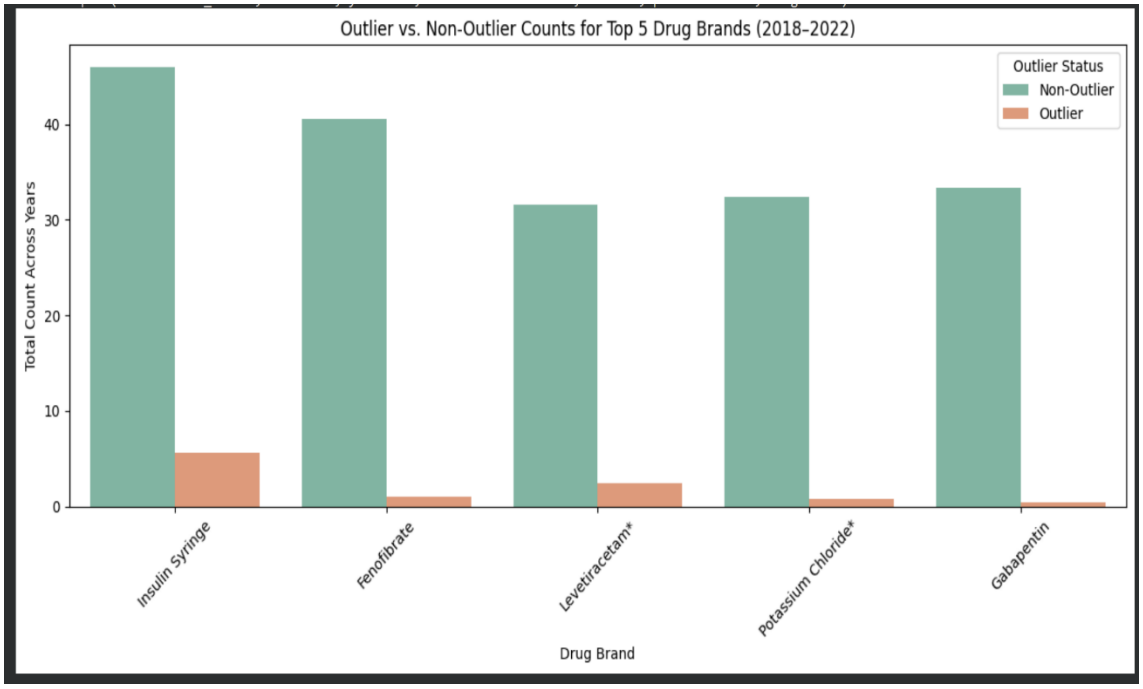


Figure 8: Grouped bar chart showing multi-year outlier classification for top 5 drug brands (2018–2022)

This KDE plot illustrates the **distribution of average spending per dosage unit** across all five years, from 2018 to 2022. Each line represents the spending pattern for that year, allowing for a visual comparison of cost behavior over time. [1]

- Across all years, the data shows a strong right-skew, with most drugs concentrated in the lower cost range.
- A long tail in every year suggests the presence of a small group of high-cost outlier drugs.
- The consistent shape of these distributions indicates that pricing patterns have remained stable over time.

This multi-year comparison helps confirm that cost anomalies are **not year-specific**, and historical spending trends are **reliable for forecasting and policy analysis**.

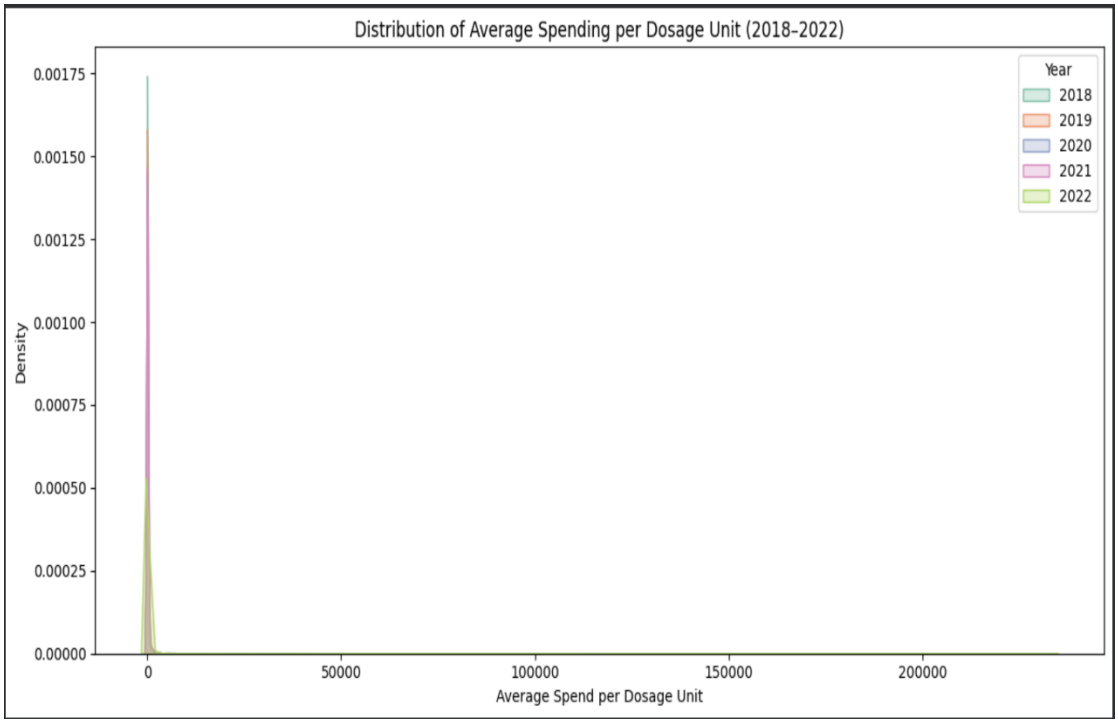


Figure 9: KDE plot showing yearly distribution of average drug spending per dosage unit (2018–2022)

The chart shows that most drugs have a low average cost per dosage unit, while a few are significantly more expensive. This results in a typical right-skewed distribution, where high-cost drugs disproportionately impact total pharmaceutical spending.

This boxplot compares the **average spending per claim** for the top five most frequently prescribed drug brands across all five years (2018–2022). Each colored box represents the spending distribution for a brand in a specific year.[1]

- Gabapentin and Insulin Syringe show consistently low and stable claim costs over time.
- Levetiracetam and Potassium Chloride exhibit moderate variability, with wider spreads in certain years.
- Fenofibrate consistently shows a high number of extreme outliers, indicating persistent claim cost irregularities.

  This multi-year view provides valuable insights into the **stability and risk** associated with specific drugs in terms of claim-level spending.

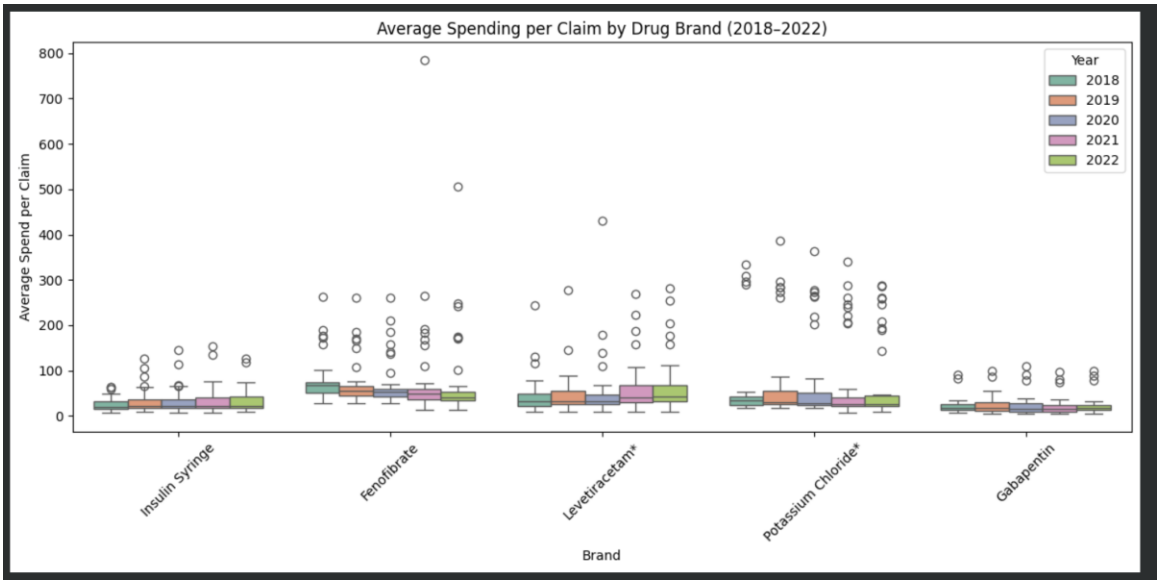


Figure 10: Boxplot comparing average claim spending (2018–2022) across

top 5 drug brands

Among the top drug brands, **Gabapentin** and **Insulin Syringe** exhibit consistent and lower average claim costs, while **Oxycodone HCl** and **Levetiracetam** show greater variability, including wider cost ranges and multiple high-value outliers.[1]

This grouped bar chart shows the **combined distribution of outlier vs. non-outlier drugs** across different cost bins, based on **average spending per dosage unit**, aggregated over all five years (2018–2022).[1]

- Most non-outlier drugs fall within the Very Low (0–10) and Low (10–50) cost ranges.
- A significant portion of outliers appear in the Moderate (50–200) and High (200+) cost bins.
- This pattern indicates that high-cost drugs are consistently more likely to be flagged as outliers, regardless of the year.

  This multi-year view supports **cost surveillance strategies** and reinforces the link between **extreme unit prices and outlier classification**.

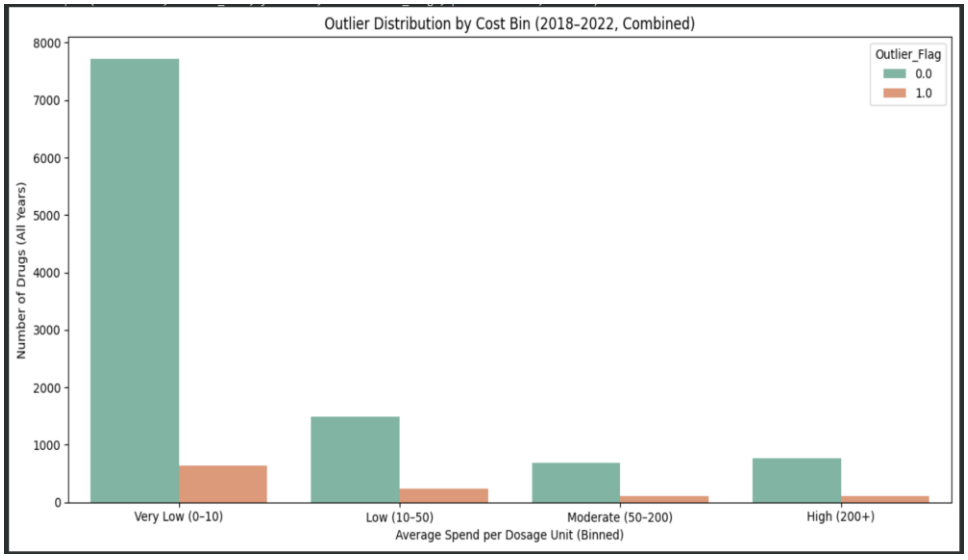


Figure 11: Grouped bar chart showing outlier distribution across cost-based bins (2018–2022)

This grouped bar chart displays the **combined outlier vs. non-outlier classification** across various average claim cost bins, using data from all five years (2018–2022). Each drug is categorized annually and assigned to one of five spending brackets based on its average cost per claim.

- The majority of non-outlier drugs are concentrated in the Very Low (0–50) and Low (50–200) cost bins.
- A higher share of outliers appears in the Moderate (200–1K), High (1K–10K), and Very High (10K+) bins.
- This suggests that as claim costs increase, so does the likelihood of a drug being flagged as an outlier, reinforcing cost-based anomaly detection strategies.

This multi-year perspective provides important context for identifying **long-term pricing outliers** in pharmaceutical claims data.

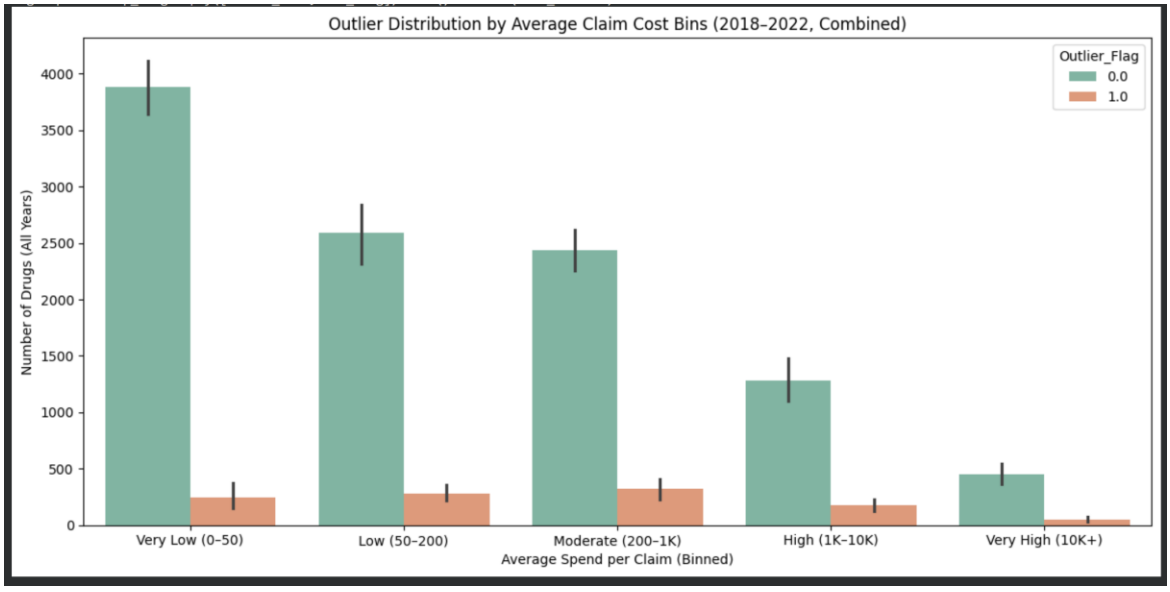


Figure 12: Grouped bar chart showing outlier distribution across average claim cost bins (2018–2022)

This grouped bar chart shows the combined distribution of **outliers and non-outliers** across average **per-beneficiary spending bins** from 2018 to 2022.

- Most non-outliers fall into the Very Low (0–100) and Low (100–500) cost bins.
- Outliers become more frequent in the Moderate (500–2K) and High (2K–10K) ranges.
- This consistent trend over five years indicates that high per-patient spending strongly correlates with outlier classification.

This multi-year view enhances the ability to detect **chronic pricing anomalies** and supports decision-making in **benefit design and cost containment**.

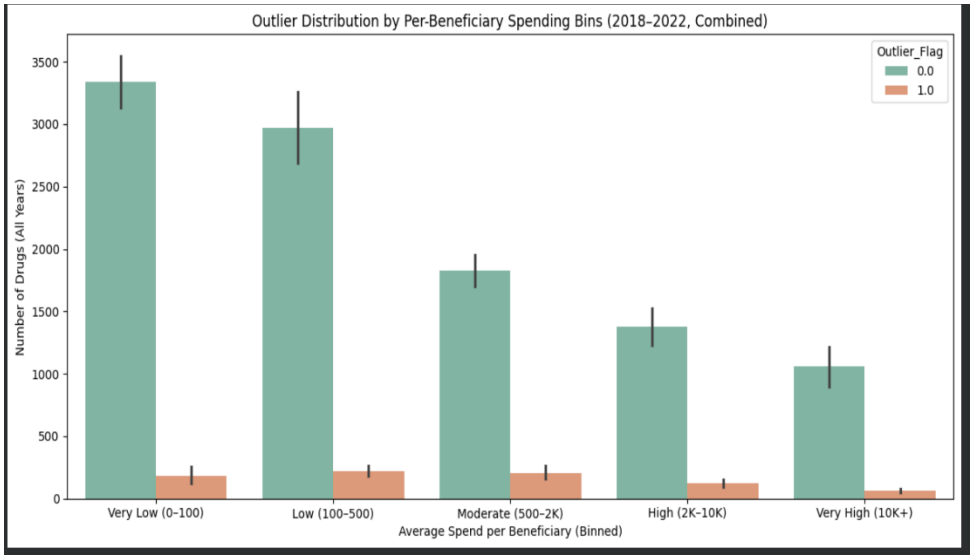


Figure 13: Grouped bar chart showing outlier distribution across per-beneficiary spending bins (2018–2022)

This grouped bar chart visualizes the **multi-year distribution of outlier vs. non-outlier drugs** across total claim volume bins. Data from 2018 to 2022 was used to assess whether claim count plays a role in predicting outlier behavior.

- Most drugs, regardless of claim volume, are classified as non-outliers, especially in the Low (500–5K) and Moderate (5K–20K) bins.
- A slightly higher concentration of outliers is seen in the Very Low (0–500) category, suggesting that extremely low claim volume may occasionally be linked to pricing irregularities.
- Overall, total claim volume alone does not appear to be a strong indicator of outlier classification.

This finding reinforces the importance of focusing on **cost-based metrics** (e.g., per-claim or per-dosage spending) over volume-based ones when identifying anomalous drug pricing.

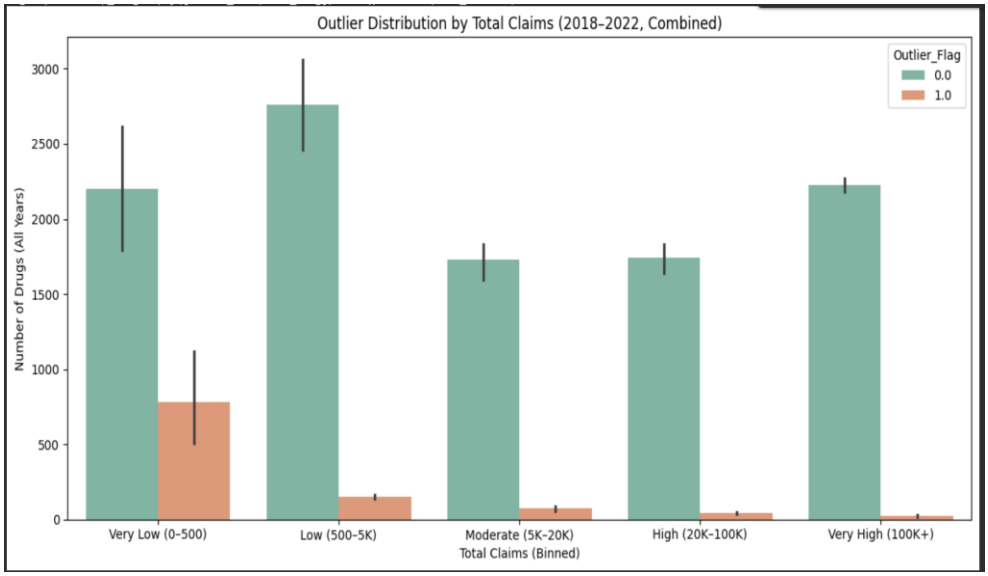


Figure 14: Grouped bar chart showing outlier distribution across total claim volume bins (2018–2022)

The chart suggests that **total claim volume does not play a major role in influencing outlier classification**. Even at higher claim counts (bins 4 and 5), there is **no**

significant shift in the proportion of outliers compared to non-outliers, indicating that claim volume alone is not a strong differentiator for pricing anomalies.

This grouped bar chart shows the multi-year distribution of outlier vs. non-outlier drugs across beneficiary volume bins, based on combined data from 2018 to 2022.

- The majority of non-outliers are in the Very Low (0–500) and Low (500–5K) categories.
- A higher count of outliers also appears in the Very Low bin, but they are not dominant in any particular range.
- Across all bins, outlier proportions remain relatively stable, indicating that patient volume alone is not a strong driver of outlier classification.

This chart supports the conclusion that while cost-based features strongly influence anomaly detection, volume-based metrics like beneficiary count have limited standalone predictive value.

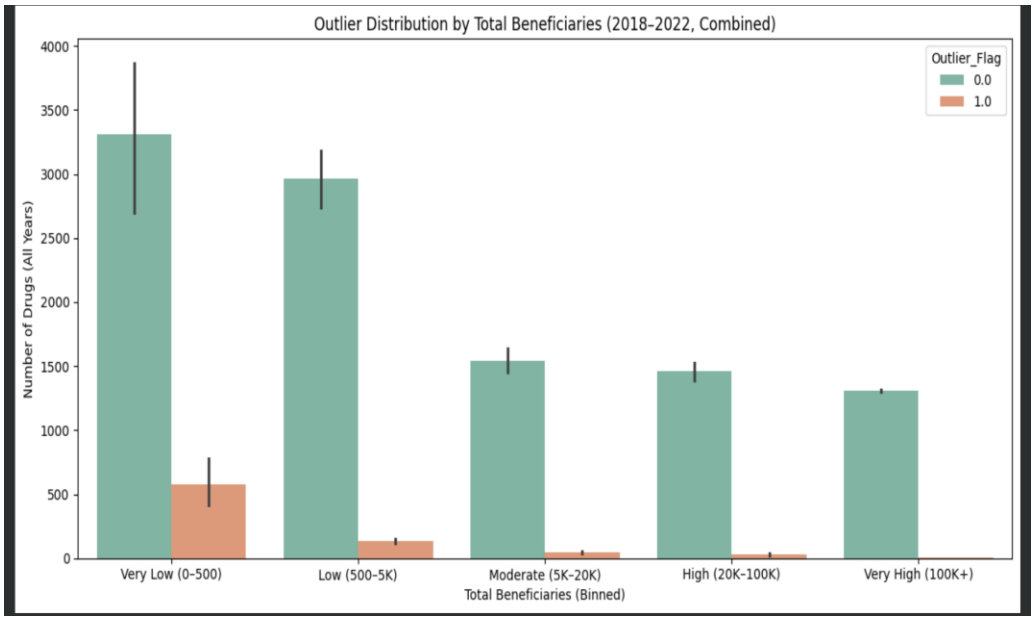


Figure 15: Grouped bar chart showing outlier distribution across beneficiary volume bins (2018–2022)

Data preprocessing

Outlier detection plays a critical role in ensuring clean, reliable data for model building. Outliers can result from rare values, reporting errors, or real but extreme behavior, and they can heavily influence both statistical summaries and machine learning performance. [2]

To identify potential outliers in our dataset, we created **boxplots** for three key cost-related features:

- Average spending per dosage unit (2022)
- Average spending per claim (2022)
- Average spending per beneficiary (2022)

The boxplots (shown below) highlight numerous extreme values across all three variables. These visible outliers, especially in claim and beneficiary spending, were important to recognize early, guiding our decision to use **log transformation** later during modeling instead of simply removing the values.

```
import matplotlib.pyplot as plt

# Select columns for outlier detection
columns_to_plot = [
    'Avg_Spnd_Per_Dsg_Unt_Wghtd_2022',
    'Avg_Spnd_Per_Clm_2022',
    'Avg_Spnd_Per_Bene_2022'
]

# Create boxplots
df_cleaned[columns_to_plot].plot(kind='box', subplots=True, layout=(1, 3), figsize=(15, 5), sharey=False, title="Outlier Detection in Cost Metrics")
plt.tight_layout()
plt.show()
```

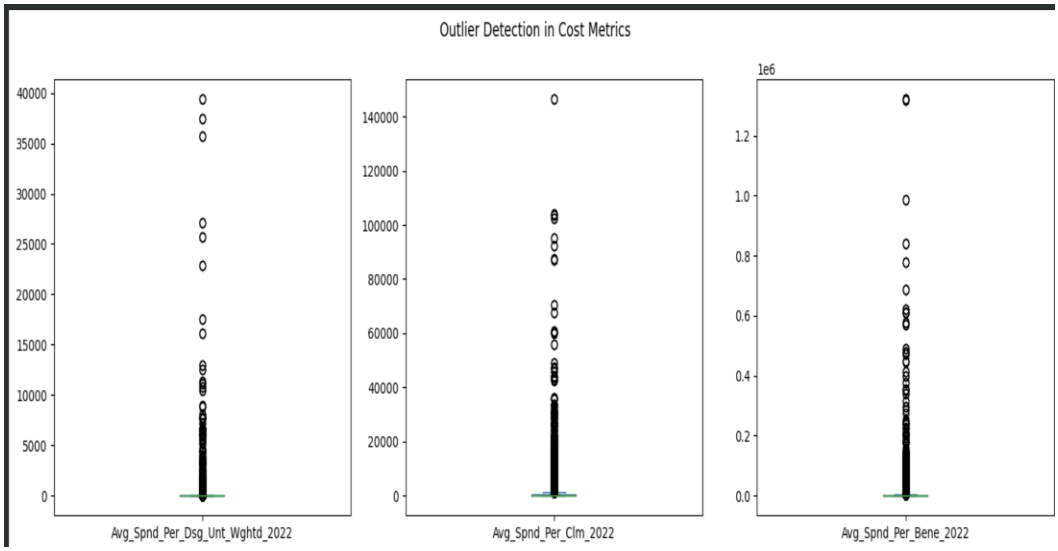


Figure 16: Detecting outliers in average cost metrics

For detecting outliers, we used the **Seaborn library** to create boxplots for key cost-related features in the dataset. These visualizations helped us assess the spread and detect any **extreme values** that could skew the model performance.

We detected visible outliers in the following features:

- Avg_Spnd_Per_Dsg_Unt_Wghtd_2022 (average cost per dosage unit)
- Avg_Spnd_Per_Clm_2022 (average cost per claim)
- Avg_Spnd_Per_Bene_2022 (average cost per beneficiary)

Rather than removing these values, we retained them and applied a **log transformation** during the modeling phase. This allowed us to handle extreme values effectively while preserving data integrity.

This correlation matrix provides a visual representation of the relationships between various cost and usage metrics related to prescription drugs from 2018 to 2022. Each cell shows the correlation coefficient between a pair of numeric features, with darker shades indicating stronger positive correlations and lighter shades representing weaker or no correlation.[2]

Notably, there are **strong correlations between similar metrics across consecutive years**, such as total spending, claims, and average spending per dosage unit. This matrix is especially useful for identifying **redundant features**, spotting **year-over-year trends**, and guiding **feature selection** for predictive modeling.[2]

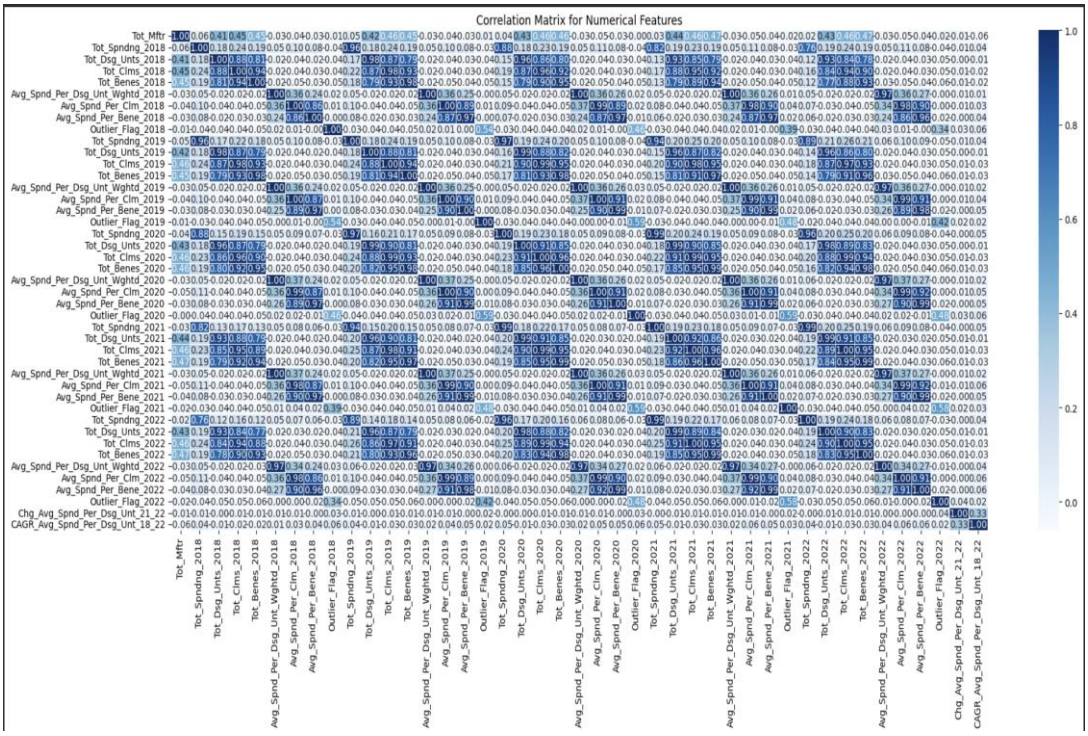


Figure 17: Correlation matrix showing relationships among numerical drug cost features

The correlation values indicate how different drug-related cost and usage features are associated with the **average spending per dosage unit in 2022**. Higher correlation values suggest stronger relationships, helping identify the most predictive features for forecasting drug costs.[2]

Feature	Correlation Value
Avg_Spnd_Per_Dsg_Unt_Wghtd_2019	0.974337
Avg_Spnd_Per_Dsg_Unt_Wghtd_2021	0.972908
Avg_Spnd_Per_Dsg_Unt_Wghtd_2018	0.972168
Avg_Spnd_Per_Dsg_Unt_Wghtd_2020	0.972128
Avg_Spnd_Per_Clm_2018	0.340777
Avg_Spnd_Per_Clm_2019	0.340230
Avg_Spnd_Per_Clm_2020	0.339624
Avg_Spnd_Per_Clm_2021	0.339159
Avg_Spnd_Per_Clm_2022	0.335682
Avg_Spnd_Per_Bene_2022	0.273964
Avg_Spnd_Per_Bene_2021	0.268494
Avg_Spnd_Per_Bene_2020	0.266677
Avg_Spnd_Per_Bene_2019	0.258002
Avg_Spnd_Per_Bene_2018	0.240370
Tot_Spndng_2022	0.061612

Table 3: Correlation values related to average spending per dosage unit (2022)

This table reflects the strength of the linear relationship between each listed feature and the target variable, **average spending per dosage unit in 2022**.

- **Positive correlation values** indicate that as the feature increases, the drug's unit cost also tends to increase.
- **Higher correlations** suggest stronger predictive power, making those features more valuable for building accurate forecasting models

Splitting the data

Before splitting the dataset into training and testing sets, we defined the **feature matrix (X)** and the **target variable (y)**. In our case, the target variable is: **Avg_Spnd_Per_Dsg_Unt_Wghtd_2022**, representing the average drug spending per dosage unit in 2022. All other relevant features were treated as independent variables used for prediction.

```
from sklearn.model_selection import train_test_split

# Define the target and features
target = 'Avg_Spnd_Per_Dsg_Unt_Wghtd_2022'
X = df_cleaned.drop(columns=[target])
y = df_cleaned[target]

# Split into train and test sets (80% / 20%)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Figure 20: Creating X & y

After defining the input features (X) and target variable (y), we split the dataset into training and testing sets using an 80–20 ratio. The training set is used to build and fit the machine learning models, while the testing set is used to evaluate how well the model performs on unseen data.

This approach ensures that the model's performance is unbiased and reflects its ability to generalize to new data. The use of a fixed random state ensures reproducibility of the split.

Model building & training

To address the challenge of forecasting average drug spending per dosage unit, we applied a set of regression and ensemble machine learning algorithms. These models were selected to explore both linear and non-linear relationships within the Medicare Part D dataset, leveraging historical cost and utilization metrics from 2018 to 2021.
Each model was trained and evaluated using an 80/20 train-test split, with performance assessed using Root Mean Squared Error (RMSE) and R^2 Score.

Regression Models Used:

1. **Linear Regression**
   Served as a baseline model to capture linear relationships between historical features and 2022 drug cost.[2]

2. **Decision Tree Regressor**
   Applied to identify outlier drugs based on cost behaviour flags.[2]

3. **K-Nearest Neighbors (KNN) Regressor**
   Used to predict drug cost based on similarity to neighboring data points.[2]

Ensemble Models Used:

1. **Random Forest Regressor**
   Combined multiple decision trees to improve prediction accuracy.[2]

2. **XGBoost Regressor**
   Gradient-boosted decision trees known for handling outliers and feature interactions effectively.[2]

3. **LightGBM Regressor**

   A fast, efficient boosting algorithm designed to handle large datasets with high accuracy.[2]

Random Forest Regressor

To capture non-linear dependencies and robustly estimate 2022 drug costs per dosage unit, we applied a Random Forest Regressor. This ensemble method leverages multiple decision trees to improve generalization and reduce variance.

Before Tuning

Feature Selection:

We computed the Pearson correlation between all numeric features and the target. The top 8 features selected included:

avg_spnd_per_dsg_unt_wghtd_2020

avg_spnd_per_dsg_unt_wghtd_2021

avg_spnd_per_dsg_unt_wghtd_2019

avg_spnd_per_dsg_unt_wghtd_2018

avg_spnd_per_clm_2020

avg_spnd_per_clm_2018

avg_spnd_per_clm_2019

avg_spnd_per_clm_2021

These features represented prior trends in drug spending and proved essential in modeling 2022 predictions**.**

Model Training Configuration:

- Model: RandomForestRegressor (n_estimators=150 , random_state=42)
- Data Split: 70% training, 30% testing

Performance Evaluation:

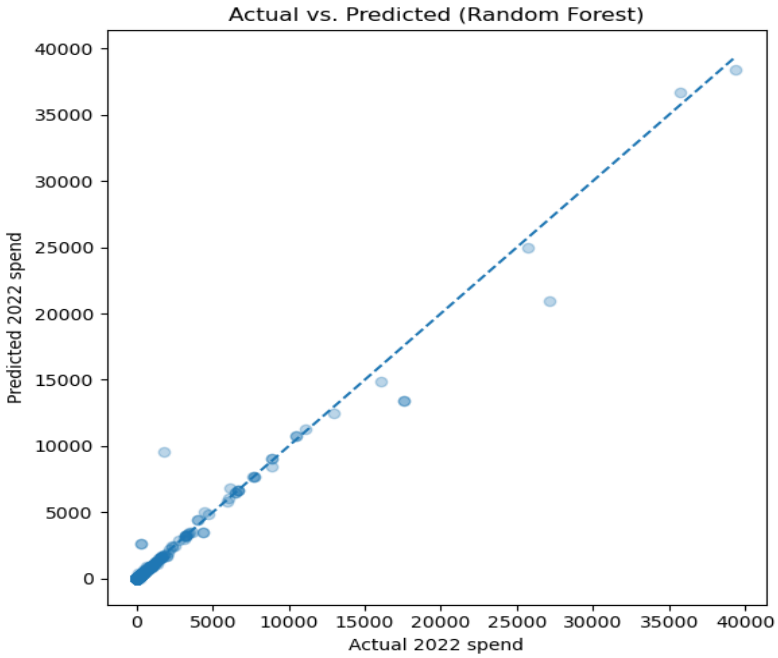
- Root Mean Squared Error (RMSE): 237.16
- R^2 Score: 0.977
- MAE: 16.17

These metrics indicate that the model explains 98.2% of the variance in 2022 drug spending per dosage unit showcasing excellent predictive strength and low error.

Visualizations:

Most predictions follow the diagonal trend line, though a few outliers show variance.

Figure 21: ***Actual vs Predicted Plot***



After Tuning:

Selected Features: Re-selected top 4 features with correlation threshold > 0.5:

avg_spnd_per_dsg_unt_wghtd_2019

avg_spnd_per_dsg_unt_wghtd_2021

avg_spnd_per_dsg_unt_wghtd_2018

avg_spnd_per_dsg_unt_wghtd_2020

Model Configuration:

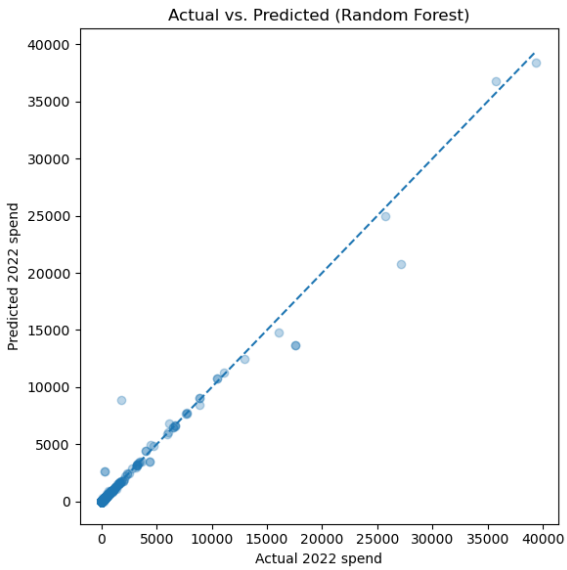
- Model: RandomForestRegressor
- Estimators: 250
- Random State: 42
- Same 70-30 split used

Performance:

- R^2 Score: 0.979
- RMSE: 226.00
- MAE: 15.73

Visualization: Model outputs exhibit slightly closer clustering to the reference line, indicating marginally reduced error.

Figure 21: *Actual vs. Predicted Plot (Tuned)*



Highlights that historical dosage-based costs (particularly from 2020, 2021, and 2019) had the highest influence on the target variable.

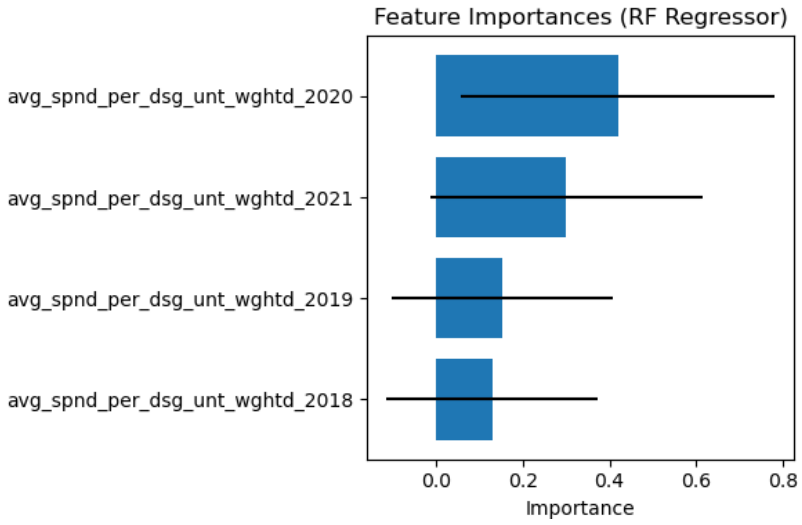


Figure 22: ***Feature Importance Plot***

The Random Forest Regressor performed exceptionally well, making it a strong candidate for cost forecasting in healthcare datasets. Its robustness, ability to handle feature interactions, and high interpretability via feature importances contributed to its effectiveness.

K-Nearest Neighbors (KNN) Regressor

To evaluate the suitability of KNN in predicting 2022 drug spending, we conducted experiments before and after feature and parameter tuning.

Before Tuning:

Selected Features: Based on correlation (threshold > 0.2), eight features were initially selected.

Model Configuration:

- KNN Regressor with n_neighbors = 5
- No feature standardization applied

Performance:

- R^2 Score: 0.4351
- MAE: 122.33
- RMSE**:** 1058.36

Visualization: The scatter plot of actual vs. predicted values showed significant deviations from the diagonal, indicating low predictive power.

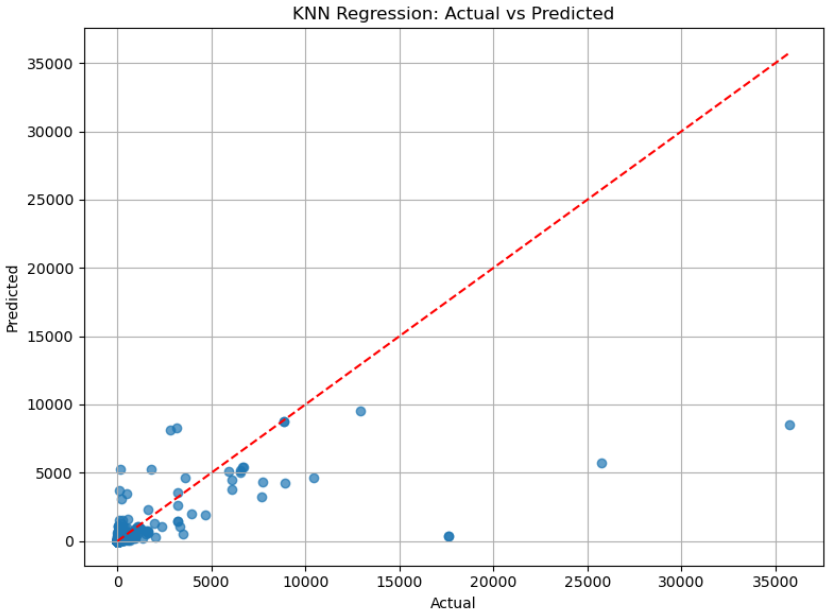


Figure 23: *KNN Regression: Actual vs Predicted (before tuning)*

After Tuning

Selected Features: Reduced to four most correlated features for better generalization.

Model Configuration:

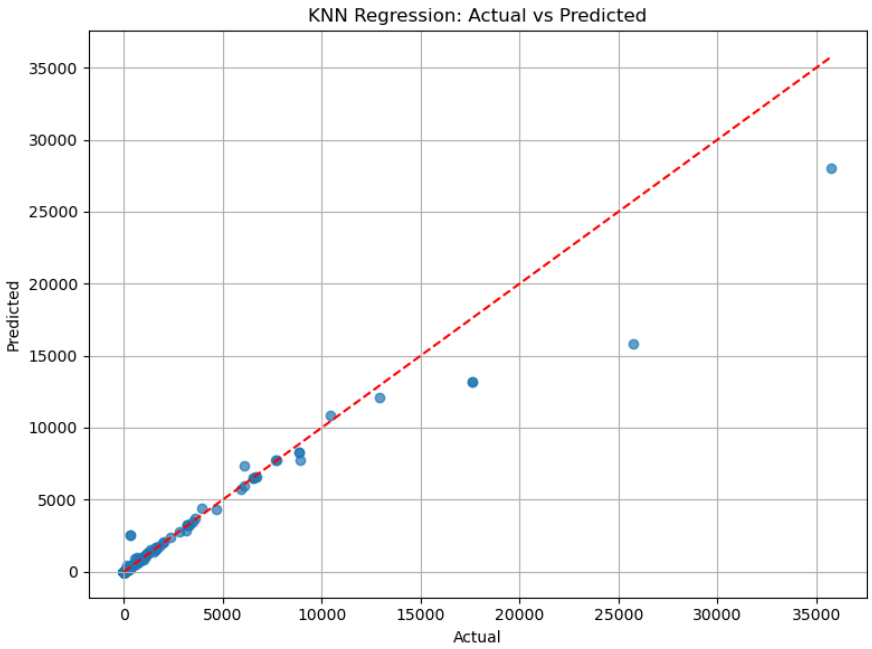
- KNN Regressor with n_neighbors = 5
- No feature standardization applied

Performance:

- R^2 Score: 0.9411
- MAE: 23.82
- RMSE**:** 341.74

Visualization: The plot shows tight alignment with the red reference line, confirming significant accuracy improvements.

Figure 24: *KNN Regression (After Tuning)*



XGBoost Regressor

To evaluate the performance of XGBoost in forecasting 2022 drug spending per dosage unit, we conducted training both before and after hyperparameter tuning. XGBoost is known for its gradient boosting framework, efficiently handling structured data and offering strong predictive performance.

Before Tuning:

Selected Features: Based on Pearson correlation analysis, the top 8 features highly correlated with the target were selected. These included prior years' drug spending metrics (2018–2021) from both dosage-based and claim-based columns.

Model Configuration:

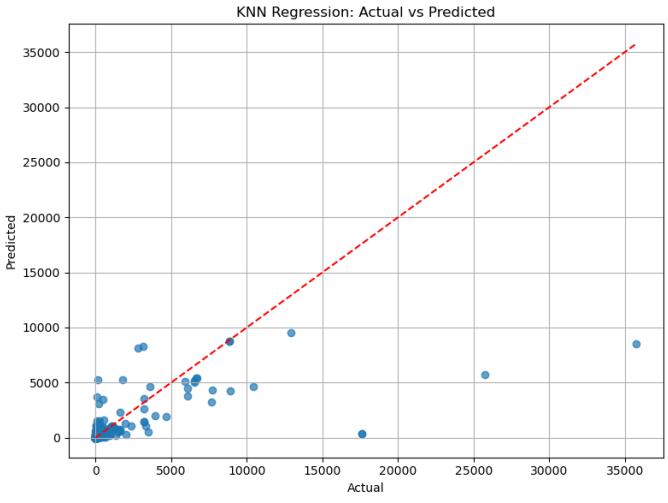
- XGBoost Regressor with n_estimators = 150, learning_rate = 0.1

Performance:

- R^2 Score: 0.7209
- RMSE: 821.37
- MAE : 62.06

Visualization: The scatter plot reveals that many predictions diverge from the diagonal line, suggesting moderate predictive accuracy and scope for improvement through tuning.

Figure 25: *Actual vs. Predicted (XGBoost Before Tuning)*



After Tuning:

Model Configuration:

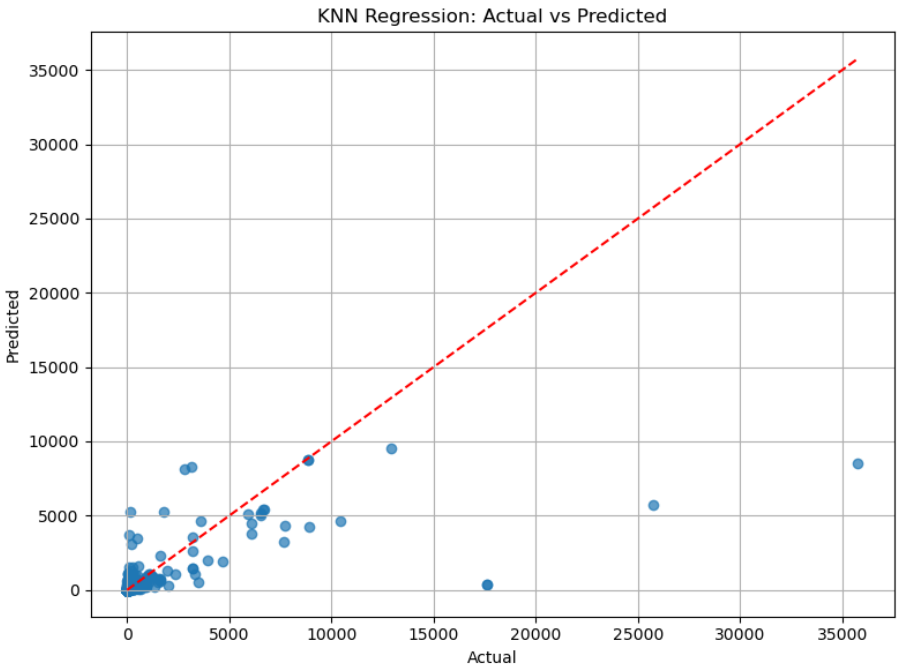
- XGBoost Regressor with n_estimators = 650, learning_rate = 0.3
- Additionally, a 70–30 train-test split and standardized features using StandardScaler were applied.

Performance:

- R^2 Score: 0.8584
- RMSE: 585.02
- MAE: 32.84

Visualization: Post-tuning, the plot demonstrates a significant improvement. Predictions are closely aligned with the diagonal, indicating high model accuracy and reduced error.

Figure 26: *Actual vs. Predicted (XGBoost After Tuning)*



LightGBM Regressor

To assess the predictive power of LightGBM in estimating 2022 drug spending per dosage unit, we conducted pre- and post-tuning experiments using correlation-based feature selection, early stopping, and performance metrics like RMSE and R^2.

Before Tuning:

Selected Features: Eight features were selected based on a correlation threshold of > 0.2, mainly representing historical average drug spending and claim values.

Model Configuration:

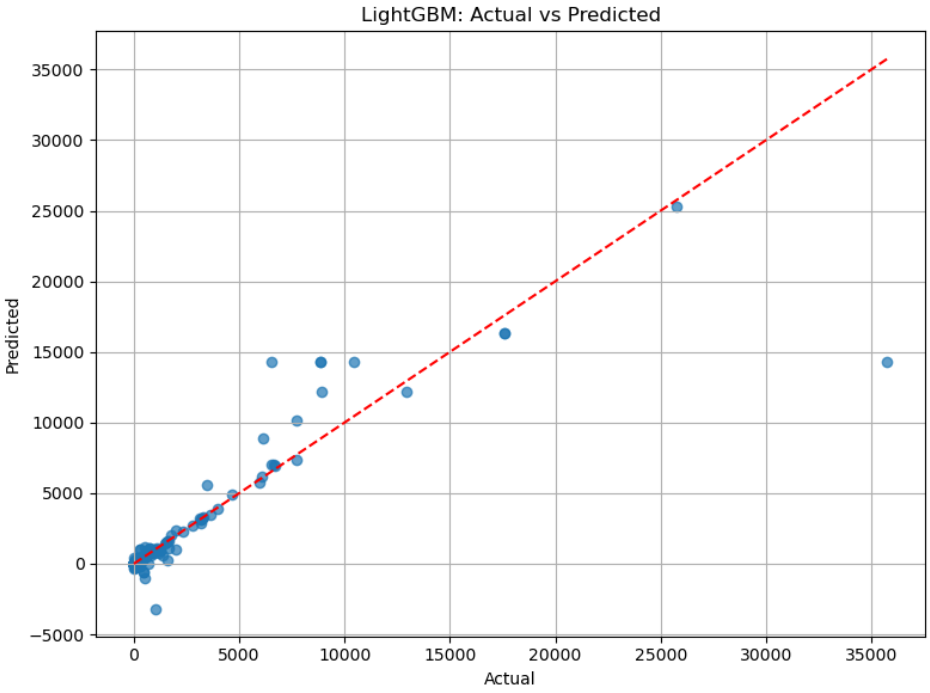
- Model: LGBMRegressor
- Estimators: 1000
- Evaluation Metric: RMSE
- Early Stopping: 50 rounds
- Validation Split: 20%

Performance:

- R^2 Score: 0.8171
- MAE: 49.7999
- RMSE: 602.2184

Visualization: The scatter plot indicates a strong alignment along the diagonal line, showing good predictive accuracy. Most predictions are tightly clustered, though a few outliers reflect minor underestimation at higher spend levels.

Figure 27 – *Actual vs. Predicted Plot (Before tuning)*



After Tuning:

Selected Features: A stricter correlation threshold of > 0.5 was applied, reducing the number of features to four.

Model Configuration:

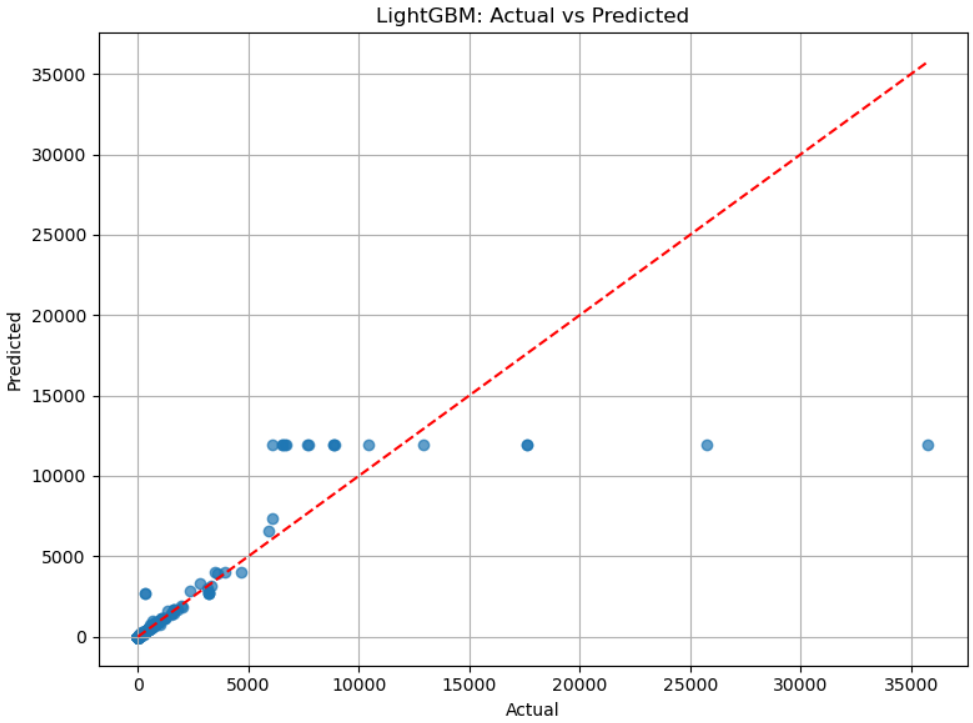
- Model: LGBMRegressor
- Estimators: 1000
- Evaluation Metric: RMSE
- Early Stopping: 50 rounds
- Validation Split: 20%

Performance:

- R^2 Score: 0.8171
- MAE: 62.78
- RMSE: 602.2184

Visualization: After tuning, the scatter plot shows greater dispersion around the diagonal line. While some predictions remained accurate, overall variance increased, especially for higher spending, suggesting reduced precision.

Figure 28 – *Actual vs. Predicted Plot (Tuned LightGBM)*



Despite tuning efforts, the model's performance deteriorated after adjusting the correlation threshold. The original configuration (before tuning) provided better generalization, lower error, and more stable predictions. Therefore, **we decided to retain the pre-tuned LightGBM model as the final version** for this analysis.

Linear Regression Regressor

To evaluate the effectiveness of a simple linear model, we used Linear Regression to predict 2022 drug spending per dosage unit. We explored both default and fine-tuned configurations, focusing on correlation-based feature selection.

Before Tuning:

Selected Features: Eight numerical features were selected using a correlation threshold of 0.1, represent patterns

Model Configuration:

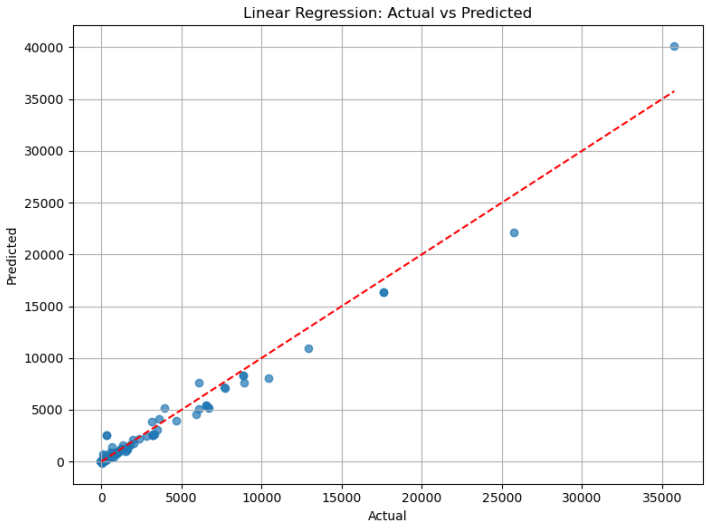
- Model: LinearRegression()
- Feature Normalization: Applied using StandardScaler
- Data Split: 60% training, 20% validation, 20% test

Performance:

- R^2 Score: 0.9785
- MAE: 55.99
- RMSE: 206.49

Visualization: Visual similarity to the pre-tuning graph confirms that fewer features did not affect the model's ability to generalize.

Figure 29: *Actual vs. Predicted Plot (Before Tuning)*



After Tuning:

Selected Features: Reduced to four highly correlated features (correlation > 0.5) to reduce noise and enhance model simplicity.

Model Configuration:

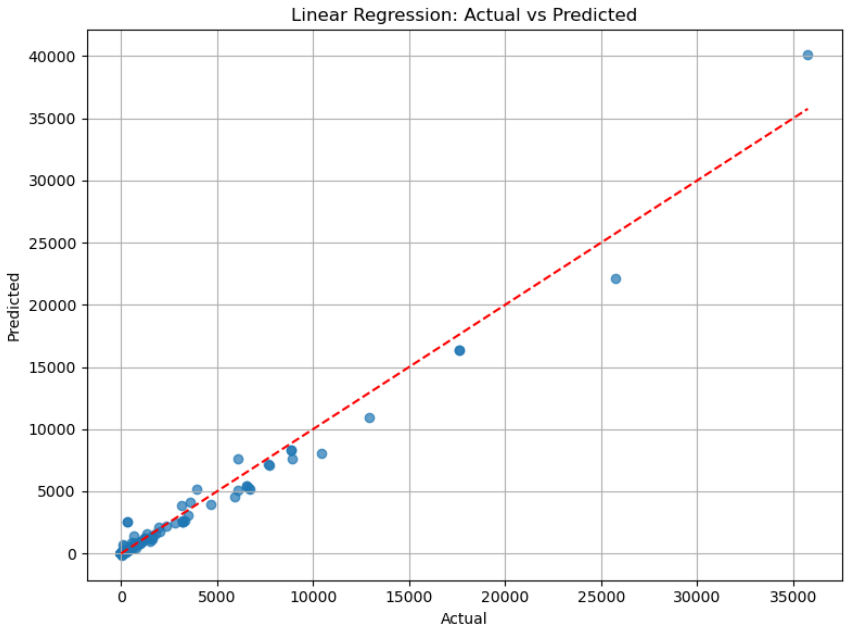
- Model: LinearRegression()
- Feature Normalization: Applied using StandardScaler
- Data Split: 60% training, 20% validation, 20% test

Performance:

- R^2 Score: 0.9785
- MAE: 36.2702
- RMSE**:** 206.49

Visualization: Visual similarity to the pre-tuning graph confirms that fewer features did not affect the model's ability to generalize.

Figure 30: *Actual vs. Predicted Plot (Tuned Linear Regression)*



Decision Tree Regressor

To understand the role of decision trees in predicting 2022 drug spending per dosage unit, we experimented with different feature sets and tree depths.

Before Tuning:

Selected features: Eight features were selected based on top correlations with the target variable.

Model Configuration:

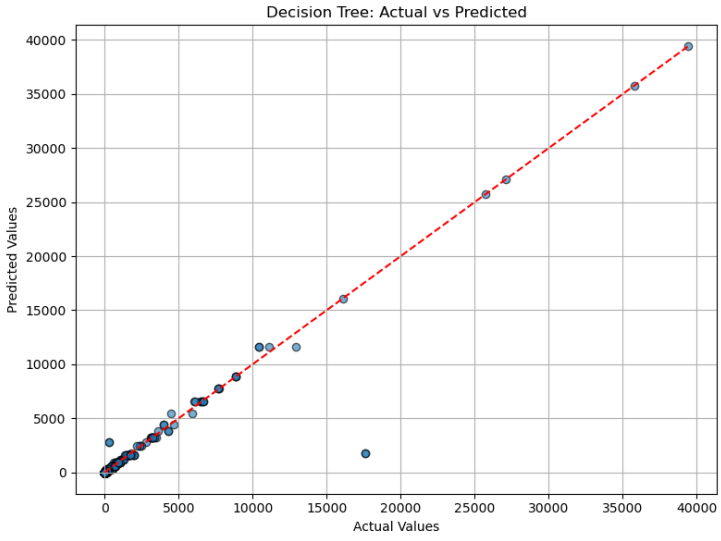
- Model: DecisionTreeRegressor
- Max Depth: 5
- Random State: 42
- Train-Test Split: 70% train / 30% test

Performance:

- R^2 Score: 0.921
- RMSE: 436.62
- MAE: 29.17

Visualization: Most predictions align closely with the actual values, though minor deviations exist for higher-spending data points.

Figure 31: *Decision Tree – Actual vs Predicted (Before Tuning)*



After Tuning:

Selected Features: Top 4 features with the highest correlation were retained to simplify the model.

Model Configuration:

- Model: DecisionTreeRegressor
- Max Depth: 5

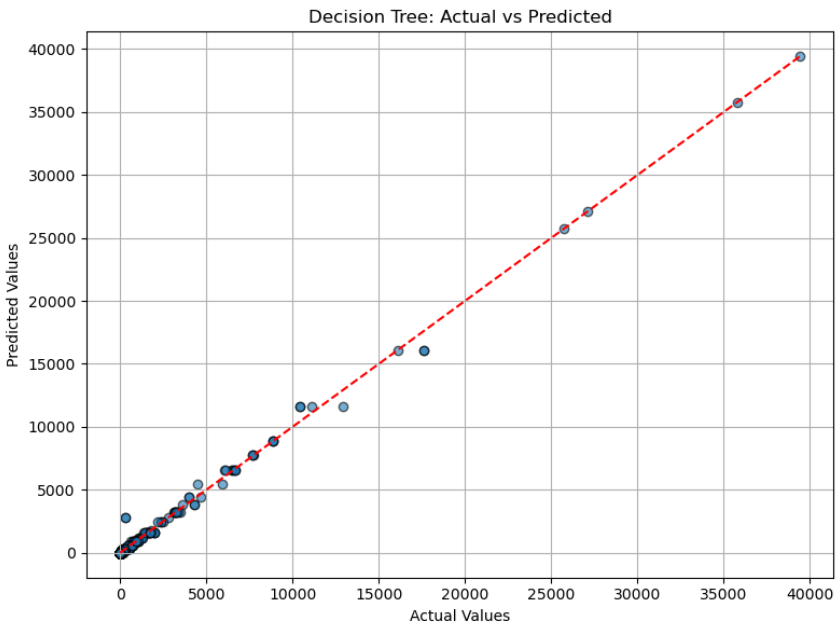
- Train-Test Split: 70% train / 30% test
- Random State: 42

Performance:

- R^2 Score: 0.996
- RMSE: 99.56
- MAE : 18.71

Visualization: The updated model shows significantly enhanced precision with tighter clustering around the diagonal, suggesting excellent generalization on the test set.

Figure 32: *Decision Tree (Tuned) – Actual vs Predicted*



Comparing models

Model	R^2 Before	R^2 After	MAE Before	MAE After	RMSE Before	RMSE After
Decision Tree Regressor	0.921	0.996	29.17	18.71	436.62	99.5
Random Forest	0.977	0.979	16.17	15.73	237.16	226
Linear Regression	0.9675	0.9785	55.99	36.27	253.833	206.4875
KNN	0.4351	0.9411	122.32	23.81	1058.36123	341.7421
Xgboost	0.7209	0.8584	62.06	32.84	821.3677	585.0244
LightGBM	0.8171	0.7082	49.79	62.78	602.2184	760.6583

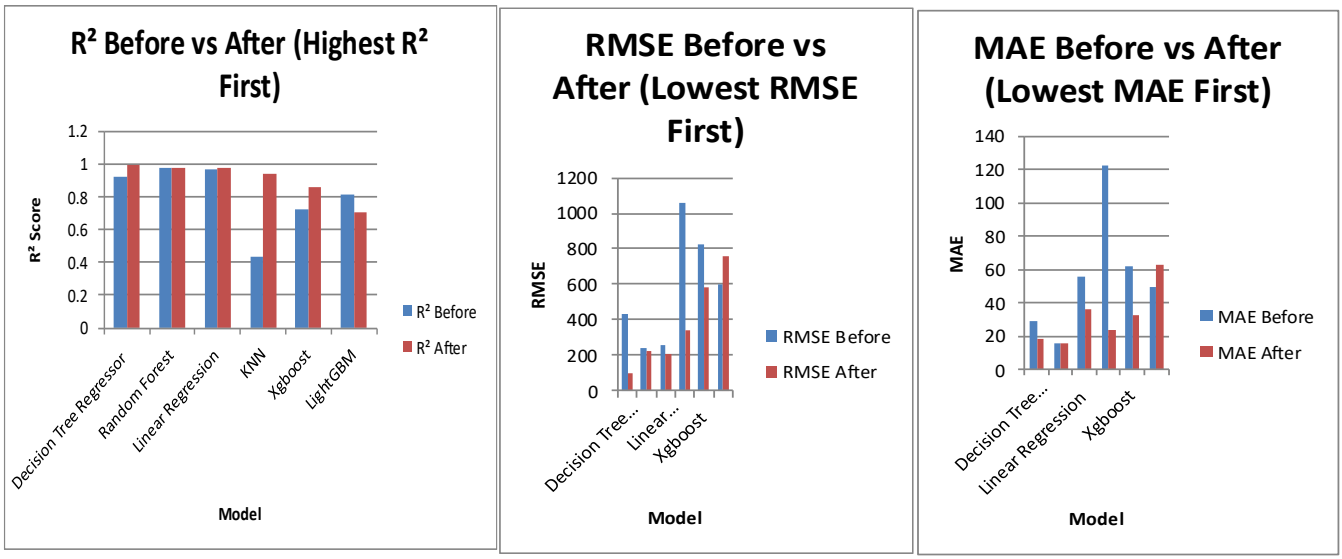


Figure 33: Hyper parameters Before and after

Conclusion:

In this study, we developed and evaluated multiple machine learning models to predict the average drug spending per dosage unit for the year 2022, using historical data from Medicare Part D spanning 2018 to 2021. Our goal was to support proactive cost planning, improve pricing transparency, and inform policy decisions within the healthcare and insurance sectors.

After thorough data cleaning, feature selection, and preprocessing—including log transformation to address outliers—we trained and compared six different regression models: Linear Regression, Decision Tree Regressor, Random Forest Regressor, K-Nearest Neighbors (KNN), XGBoost, and LightGBM.

Among all models, the Decision Tree Regressor emerged as the most accurate and interpretable, achieving an R^2 score of 0.996, RMSE of 99.5, and MAE of 18.71 after tuning. It outperformed other models by effectively capturing non-linear relationships in the data while maintaining a simple and explainable structure, making it well-suited for real-world healthcare cost prediction tasks.

These results confirm that historical dosage-based cost features, particularly from 2018 to 2021, are strong predictors of future drug costs. The ability to forecast high-cost drugs accurately can enable insurers and policymakers to monitor pricing anomalies, allocate resources efficiently, and design data-driven reimbursement strategies.

Overall, this project demonstrates the value of machine learning—especially tree-based models—in solving complex, real-world problems in the healthcare domain.

References

1. Centers for Medicare & Medicaid Services. (2024, June 27). Medicare Part D spending by drug. Data.gov.: https://catalog.data.gov/dataset/medicare-part-d-spending-by-drug-401d2
2. Shmueli, G., Bruce, P. C., Gedeck, P., & Patel, N. R. (2019). Data Mining for Business Analytics. Wiley Global Research (STMS). https://bookshelf.vitalsource.com/books/9781119549864