May 9, 2025

Predicting the Cost of Subcutaneous Tissue Debridement Across U.S. Hospitals Using Machine Learning

Agenda

1. Introduction: The Challenge of Cost Prediction
2. Motivation and Project Objectives
3. Dataset Overview
4. Data Cleaning and Preparation
5. Price Distribution and Analysis
6. Feature Engineering and Transformation
7. Modeling Approaches: Overview
8. Linear Regression: Baseline Results
9. Decision Tree Regressor: Nonlinear Modeling
10. FNN Results and Model Performance
11. Interpretability: Effects of State and Payer
12. Hyperparameter Tuning & Scaling
13. Limitations and Challenges Encountered
14. Future Directions
15. Conclusion: Key Findings and Implications
16. References

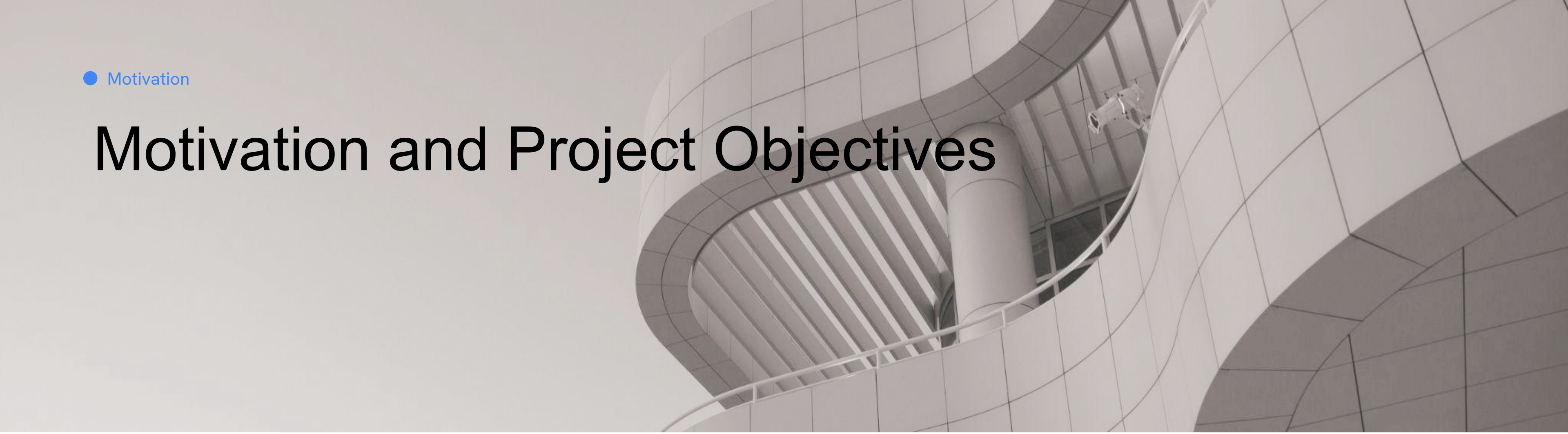
The Challenge of Cost Prediction

U.S. healthcare spending reached approximately $4.9 trillion in 2023, averaging $14,570 per person annually, far exceeding other developed nations[1].

Despite this, patients face unpredictable and often unaffordable costs due to opaque pricing and surprise out-of-network bills[3].

Policy measures like the No Surprises Act (2022) and Hospital Price Transparency Rules (2021) aim to improve cost predictability, but challenges remain in enforcement and consumer interpretation[2,4].

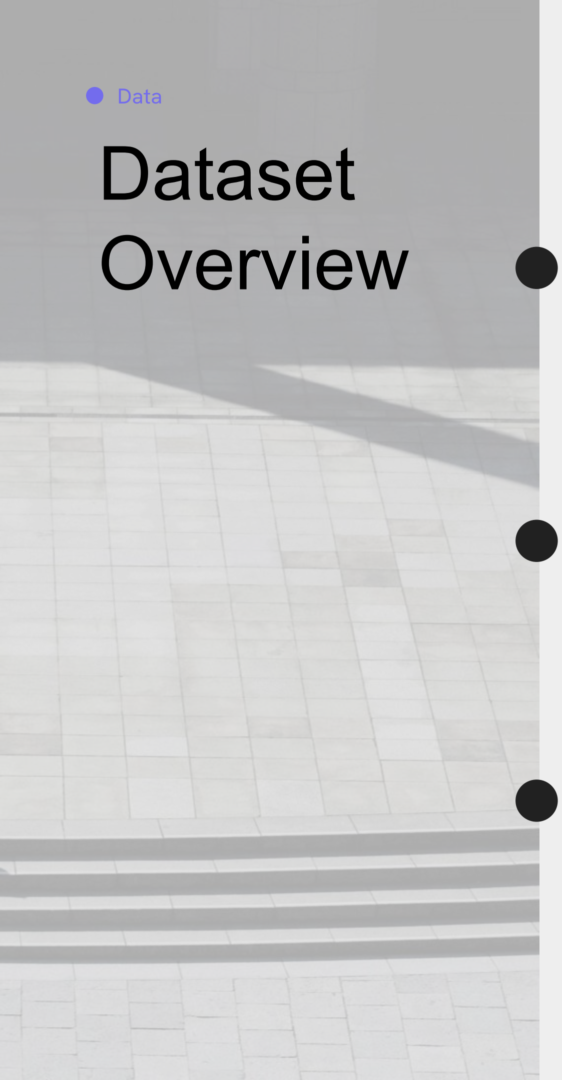
This unpredictability undermines patient trust and leads to delayed or forgone care, highlighting the urgent need for accurate cost prediction tools tailored to individual patient circumstances.



● Motivation

Motivation and Project Objectives

- Healthcare costs in the U.S. are high and often unpredictable, creating financial burdens for patients.
- Existing pricing transparency initiatives are incomplete, with many patients receiving surprise bills and lacking cost estimates before care.
- Subcutaneous tissue debridement is a niche procedure with significant price variability, making it an ideal case study for cost prediction.
- The project aims to develop a machine learning pipeline to accurately predict procedure costs using comprehensive datasets, focusing on payer type, location, and procedural metadata.
- Goals include improving price transparency, aiding patients and providers in cost planning, and uncovering regional and insurer-specific pricing dynamics.



● Data

Dataset Overview

● **CPT_HCPCS Table**

Contains procedure codes and descriptions, including CPT and HCPCS standards, which identify medical services for billing and analysis.

● **Hospital Table**

Includes hospital metadata such as unique identifiers, names, and locations, providing contextual information to link pricing data.

● **Prices Table**

Captures the core pricing data for procedures, linking hospitals and payers with procedure costs to enable cost prediction modeling.

● Preprocessing

Data Cleaning and Preparation

Initial Inspection

Prices dataset contained over 4 million records with extreme outliers; hospitals dataset had missing values in publish_date, zip_code, and street_address, and duplicated hospital names due to multiple locations.

Column Reduction

Dropped high-cardinality and inconsistent fields such as hospital name, url, street_address, city, and publish_date; removed zip_code from prices dataset to reduce noise and redundancy.

State Selection

Selected hospitals from Michigan, Florida, and Texas due to their high representation and regional diversity, ensuring a balanced and manageable dataset for analysis.

Merging and Payer Consolidation

Merged price and hospital datasets on npi_number to link pricing with state metadata; grouped low-frequency payers into an 'OTHER' category to simplify feature space and improve model robustness.

Price Distribution Analysis

- Figure 1, the initial box plot was overwhelmed by extreme outliers.
- To address this issue, we applied a trimming filter, restricting the dataset to prices less than or equal to $10,000 as shown in figure 2.
- In Figure 2, the box spans from approximately $0 to $1700. The median price falls near $700, indicating that half of the prices are below this value.
- Prices beyond $4000 are mostly outliers.

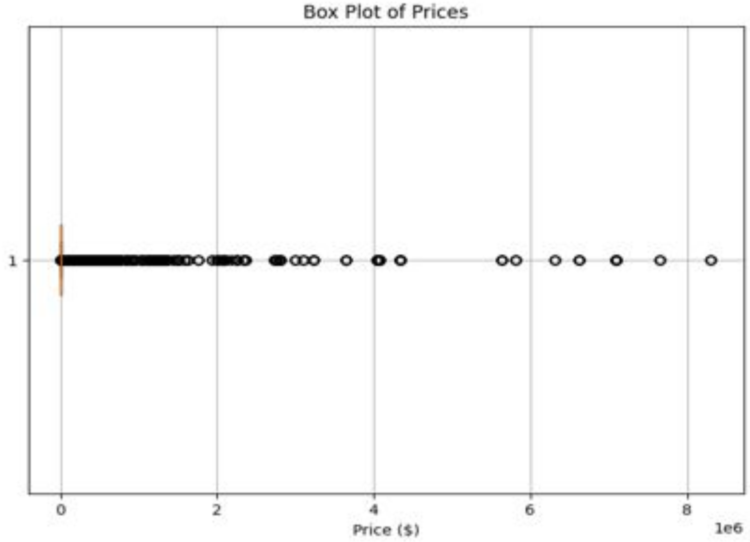


Figure 1: Box Plot before trimming prices.

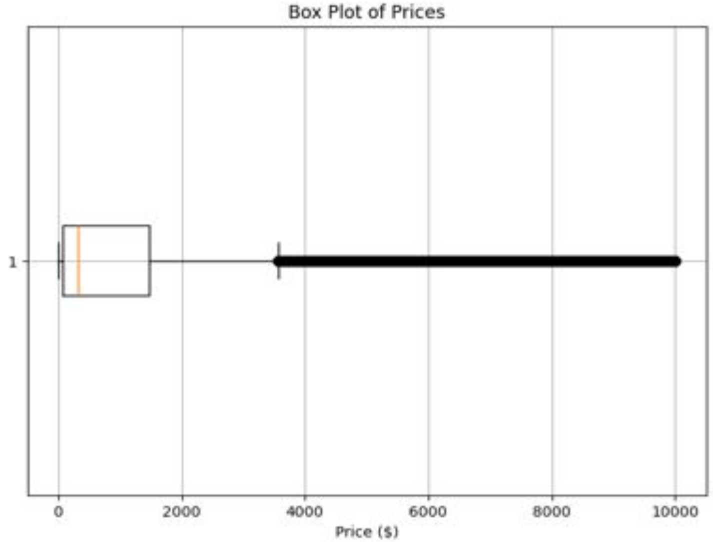
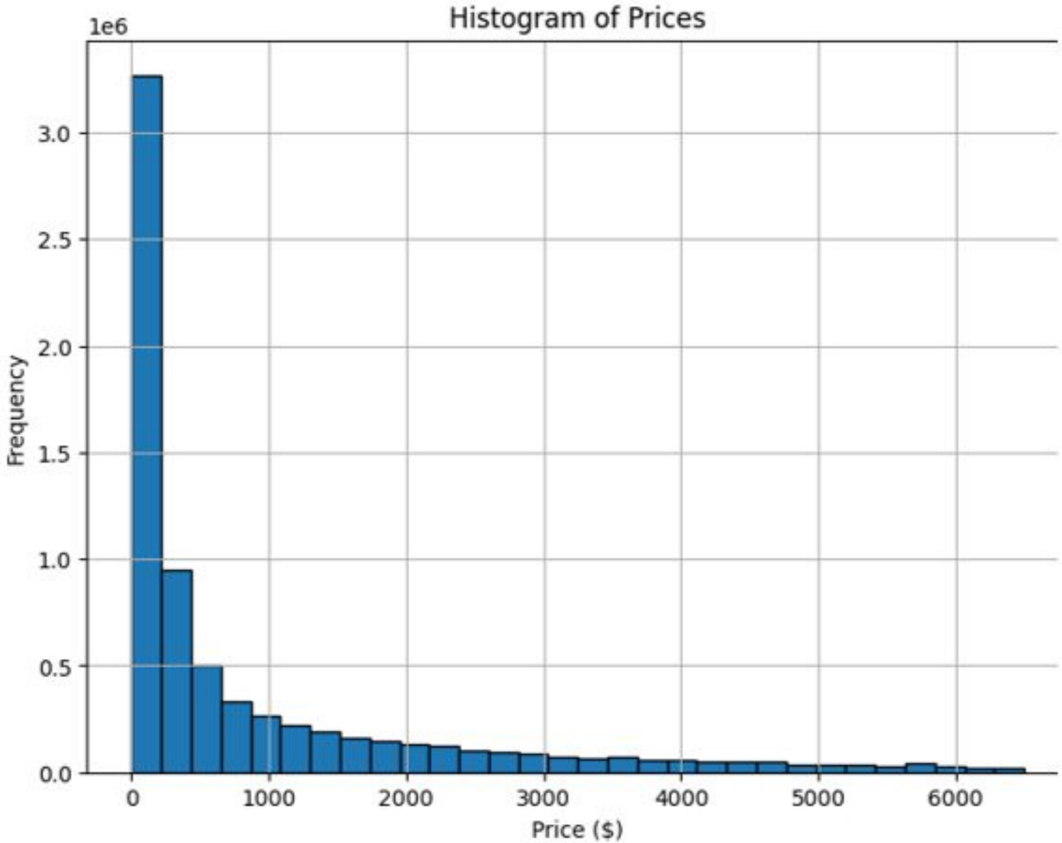


Figure 2: Box Plot after trimming prices < $10,000

Final Cleaned Dataset before merging

- To further improve interpretability, we applied an additional filter to include only prices below $6,500, narrowing the range to where the majority of values were concentrated.
- The Histogram shows a strong right skew, with the highest concentration of prices between $0 and $1000, peaking near the median.



● Methodology

Feature Engineering and Transformation

(hospital + prices) merged with cpt_hcpcs

Merged the trimmed dataset (hospital + prices) with cpt_hcpcs using procedural code.

Merged dataset contains:
- (7601117, 5)
- mean of 1200
- Std of 1900

Feature Selection + One-Hot Encoding

Selected 9 features: 1 numerical column (code) and 8 one-hot encoded payer and state variables to reduce noise and focus on relevant inputs. Dropped one payer category to avoid multicollinearity.

Categorical features encoded:
- Payers (UHC, HUMANA, AETNA, PRIORITY HEALTH)
- States (MI, FL, TX)

hospital procedure + remove redundancies

Filtered “HC REMOVAL OF DAMAGED SKIN AND UNDERLYING TISSUE” for our dataset then dropped columns like short_description.

Cleaned dataset without:
- short_description
-

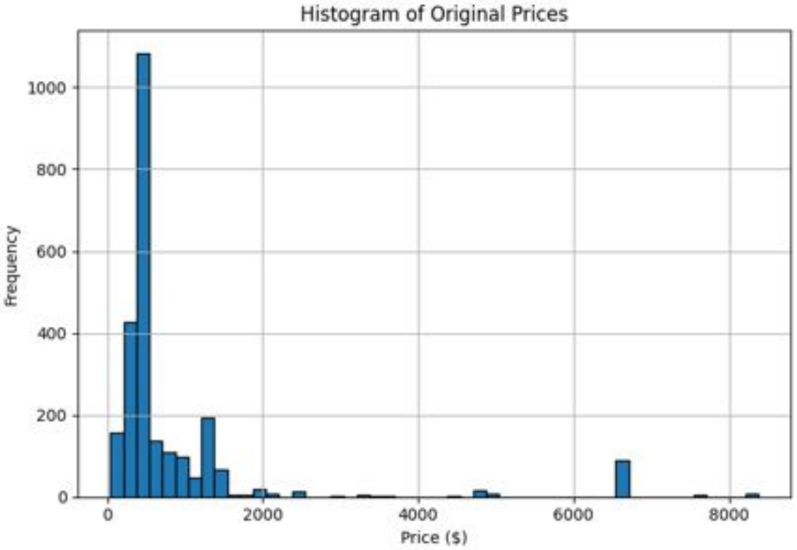
Standardized features + Log Price Transformation

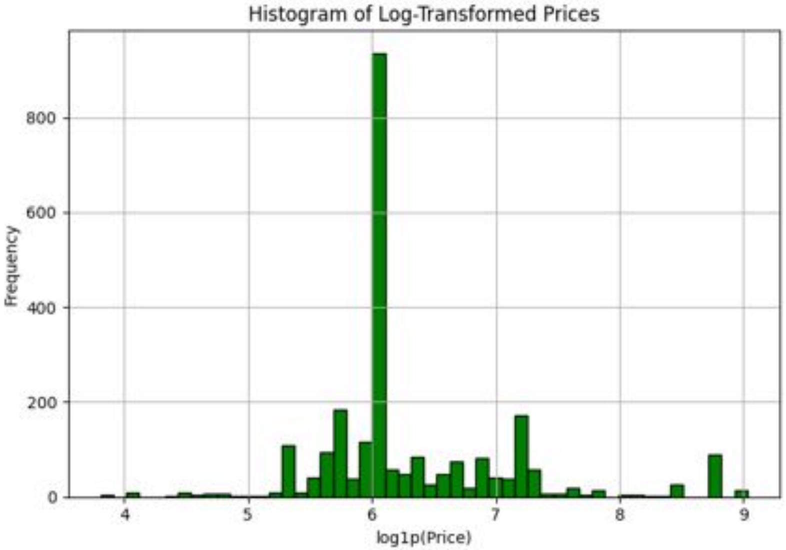
All features (X) were standardized using StandardScaler. Also applied log transformation to price using np.log1p to reduce skewness and stabilize variance for better regression.

Target variable transformed as:
- log_price = np.log1p(price)
- np.log1p(x) = log(1 + x)
- helps with small or null values of x

Log Price Transformation

We transformed the price to log scale for better model training





Evaluation Metrics

- y_i : true value of the i-th observation
- y^i : predicted value
- $\overline{y}$: mean of the true values
- n : number of observations

Mean Squared Error (MSE)

- $\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2$

Coefficient of Determination (R^2)

- $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})2}{\sum_{i=1}^{n}(y_i - \overline{y})2}$

Mean Absolute Error (MAE)

- $\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \widehat{y_i}\right|$

● Methods

Modeling Approaches: Overview

Linear Regression

Chosen as a baseline model, Linear Regression offers simplicity and interpretability by assuming a linear relationship between features and log-transformed prices, useful for initial trend analysis.

Decision Tree

Selected for its ability to model nonlinear relationships and interactions without extensive feature engineering, decision trees provide intuitive interpretability through their hierarchical splits.

Feedforward Neural Network (FNN)

Implemented to capture complex nonlinear patterns with multiple dense layers and dropout regularization, the FNN balances predictive power and generalization on healthcare cost data.

Modeling

Linear Regression: Baseline Results

Evaluation Metrics and Observations

- Linear Regression Results:
  - Real prices: R^2 = 0.72 (unitless), MSE = 370783 dollars^2, and MAE = $264.
  - Log prices: R^2 = 0.74 (unitless), MSE = 0.148 dollars^2.
- Quadratic Regression Results on real prices:
  - R^2 = 0.74 (unitless), MSE = 164849 dollars^2, and MAE = $119.

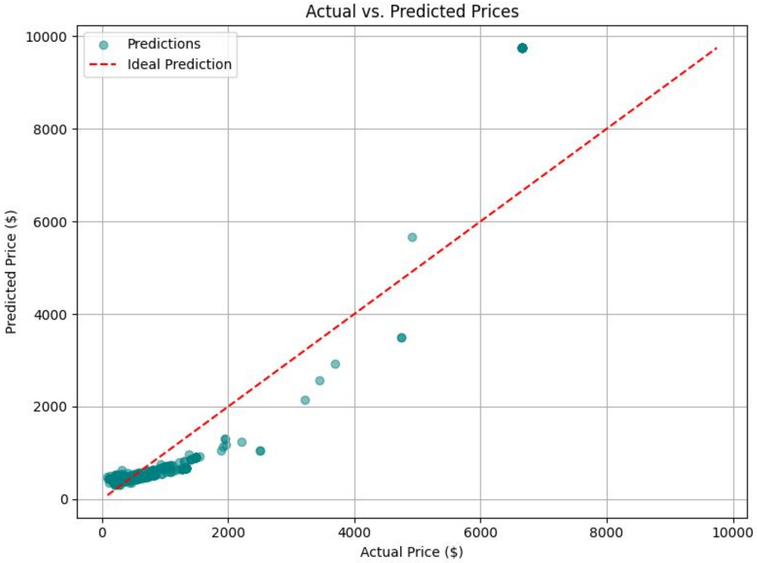


Figure 1: Linear Regression

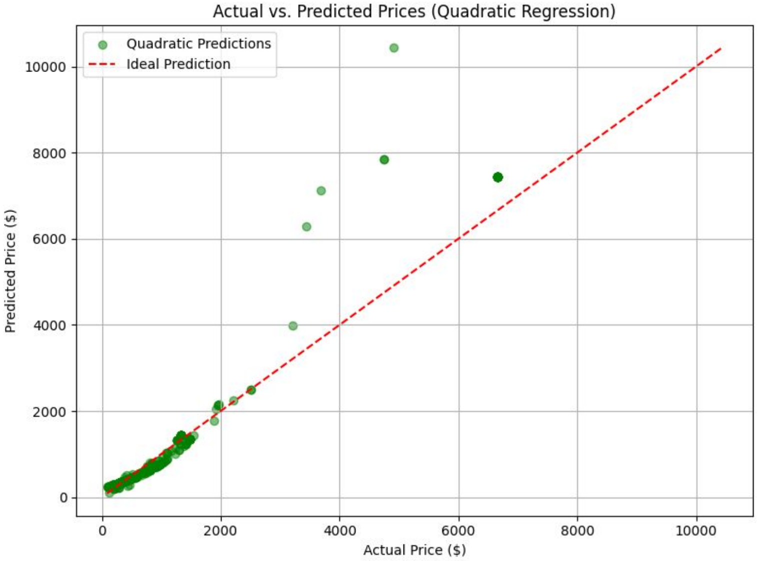


Figure 2: Quadratic Regression

Modeling

Decision Tree Regressor: Nonlinear Modeling

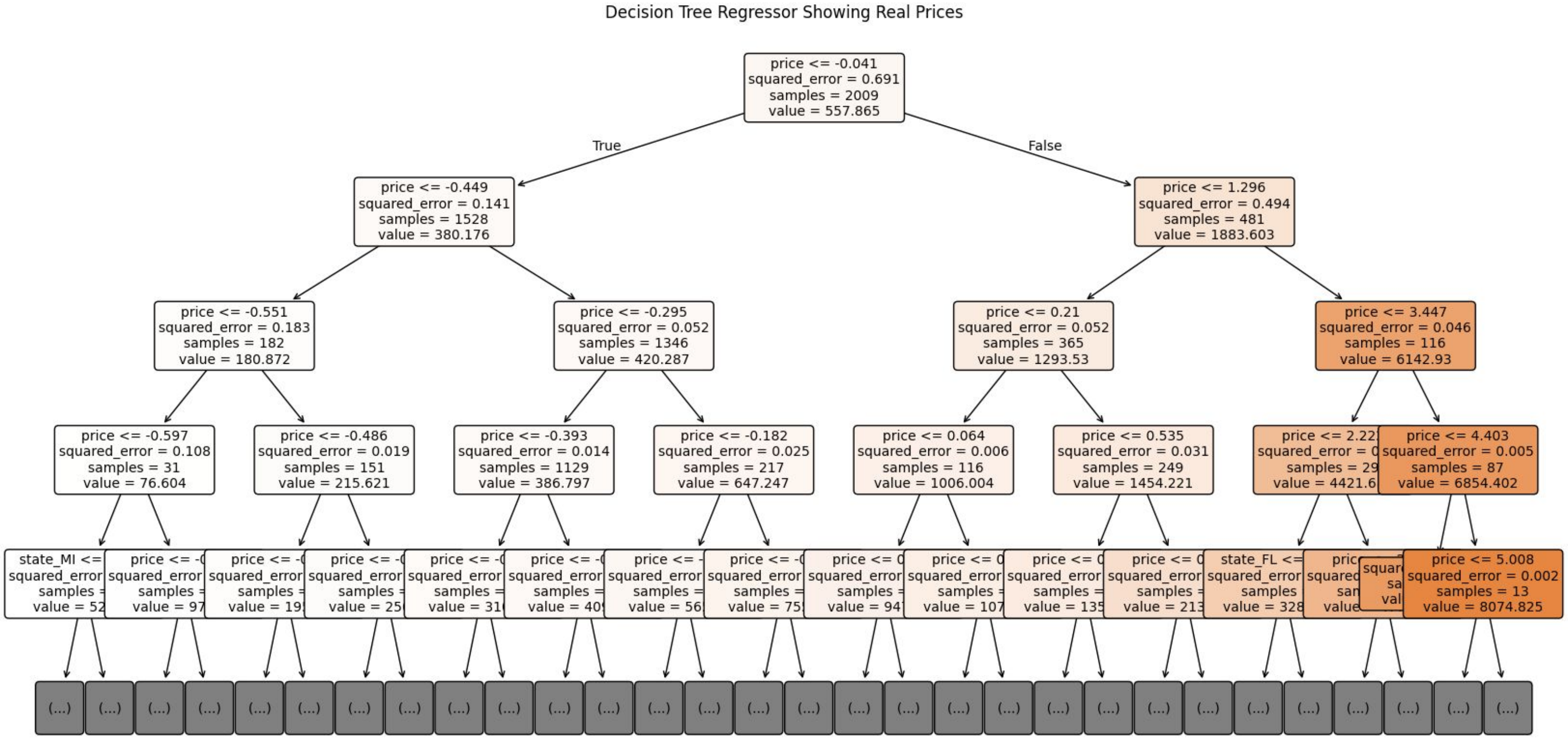
Implementation and Hyperparameters

- Used a decision tree regressor to capture nonlinear and interaction effects.
- Set max_depth to 6 to balance between accuracy and overfitting.
- Random state fixed at 42 for reproducibility.
- Criterion set to squared_error for variance reduction during splits.

Evaluation Metrics and Observations on Real Prices

- Achieved extremely high R^2 value of 0.999 (unitless), indicating near-perfect fit to training data.
- Mean Squared Error (MSE) was very low at 214.98 dollars^2, and Mean Absolute Error (MAE) was $5.01.
- Model showed clear signs of overfitting, memorizing training data patterns.
- Interpretability is strong at shallow depths but diminishes as tree complexity increases.

Decision Tree Regressor Showing Real Prices: Nonlinear Modeling

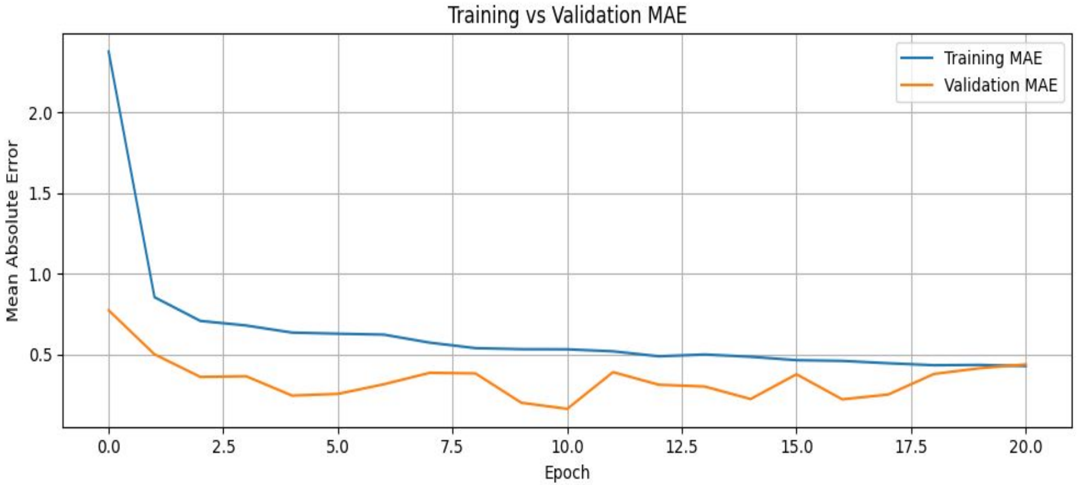


● Model Performance

FNN Results and Model Performance

Feedforward Neural Network Evaluation Metrics

- Used dense layers with Rectified Linear Unit (ReLU) as activation function
  - ReLU (x) = max(0, x)



Metrics Results on Log Prices

0.691(unitless)

R^2 Score

0.2304(dollars^2)

Mean Squared Error

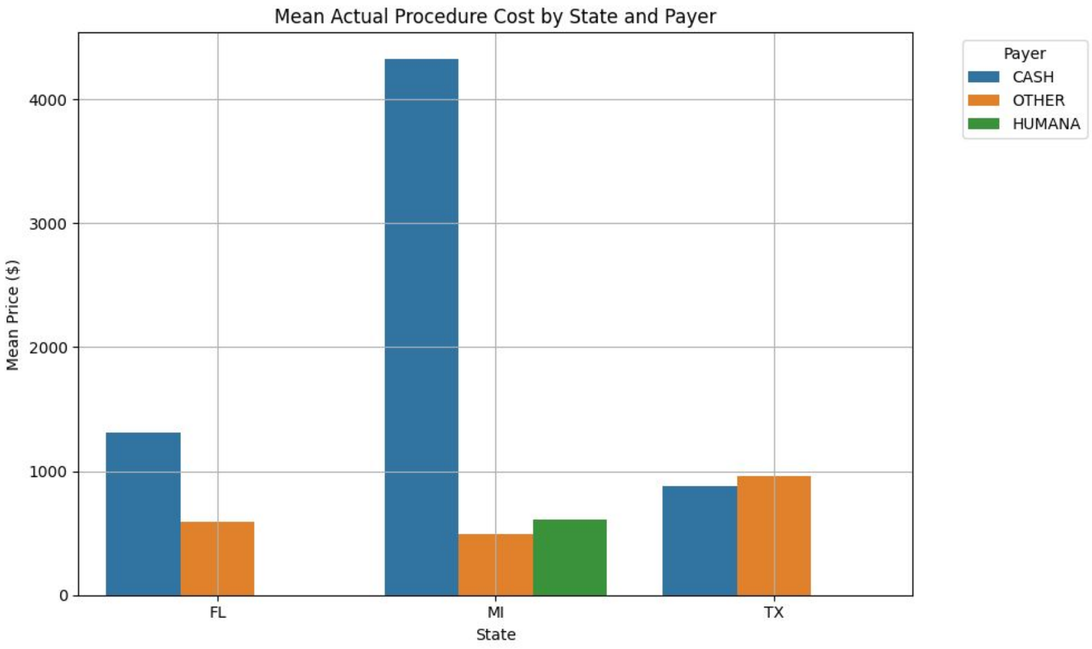
$0.4166

Mean Absolute Error

● Interpretability

Interpretability: Effects of State and Payer

Average Procedure Costs by State and Payer Group



Stratified Analysis and Insights

- Michigan exhibits the highest average procedure cost for CASH payers at $4,322, nearly three times higher than Texas, indicating marked geographic cost variation.
- Florida shows the greatest price variability among CASH payers, with a standard deviation of $944, suggesting inconsistent billing practices across hospitals.
- "OTHER" payer category is the most frequent in Michigan and Texas but consistently associated with lower average prices, possibly reflecting self-pay discounts or low-cost plans.
- Price × payer interactions, such as those involving AETNA and HUMANA, significantly affect predicted costs, underscoring insurer-specific pricing dynamics.

● Methodology

Hyperparameter Tuning & Scaling

Linear Regression

No hyperparameters were tuned; it uses ordinary least squares as a baseline model without added complexity.

Quadratic Regression

Polynomial features of degree 2 captured nonlinearities and interactions. Bias excluded to avoid duplication with linear terms, enhancing expressiveness.

Decision Tree Regressor

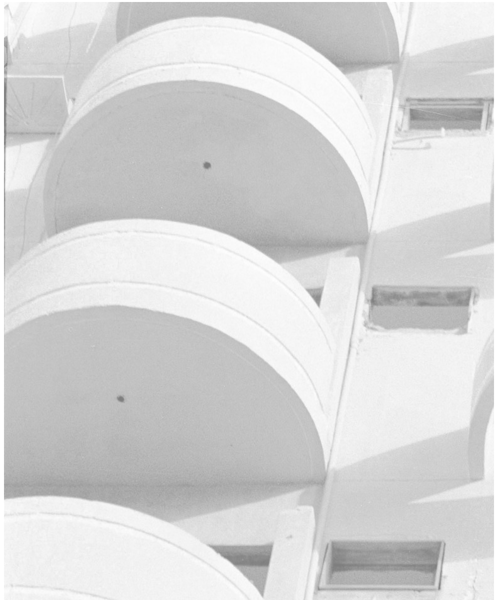
Max depth set to 6 to balance interpretability and accuracy, preventing overfitting. Random state fixed at 42 for reproducibility; squared error criterion used.

Feedforward Neural Network

Used dense layers with ReLU and 30% dropout to reduce overfitting. Early stopping with patience 10 restored best weights. Adam optimizer, MSE loss, batch size 32, max 100 epochs.

● Challenges

Limitations and Challenges Encountered



Challenges Encountered & Decisions Made

- ✓ Strong right-skew in price data required log transformation to stabilize variance and improve model fit.
- ✓ Linear regression and neural networks were trained on log-transformed prices to improve learning and reduce outlier influence.
- ✓ Decision trees performed well on real prices because they naturally handle non-linear relationships and are less sensitive to skewed distributions without needing transformation.
- ✓ Decision tree model captured complex patterns but showed clear signs of overfitting, limiting generalization.
- ✓ Grouping rare payers into an 'OTHER' category simplified modeling and reduced noise.
- ✓ Neural network architecture and dropout layers effectively reduced overfitting compared to decision trees.
- ✓ Log transformation and feature encoding enhanced compatibility and performance across models.

Impact and Limitations

- ✕ Price data included extreme outliers with values reaching millions, complicating model training and interpretation.
- ✕ Decision tree's near-perfect accuracy indicated memorization rather than generalization, reducing real-world applicability.
- ✕ Grouping diverse rare payers into a single category may obscure nuanced payer-specific cost behaviors.
- ✕ Neural networks, while better generalizing, still underpredicted very high procedure costs. They also only performed well on log prices but not when evaluated on real prices.
- ✕ Complex models like neural networks sacrifice interpretability, challenging insight extraction for stakeholders.

Conclusion: Key Findings and Implications

We developed and evaluated multiple models to predict subcutaneous tissue debridement costs using hospital pricing data. The neural network achieved the strongest generalization ($R^2 \approx 0.87$) compared to linear and quadratic models, while decision trees showed overfitting despite near-perfect accuracy.

Significant geographic and insurer-driven cost disparities were observed: Michigan exhibited the highest average cash prices, Florida showed the greatest price variability, and the "OTHER" payer category consistently had lower prices.

These findings demonstrate the feasibility and value of machine learning to uncover nuanced pricing behaviors, supporting efforts toward more transparent and equitable healthcare cost forecasting.

● Future

Future Directions

- Expand the model to cover multiple medical procedures beyond subcutaneous tissue debridement to increase practical applicability and generalizability.
- Implement clustering or embedding techniques to represent rare payer categories more effectively, preserving nuanced payer behavior while reducing dimensionality.
- Incorporate advanced machine learning methods such as ensemble models (Random Forests, Gradient Boosted Trees) and hyperparameter tuning to improve predictive accuracy and reduce overfitting.
- Enhance model transparency and fairness by integrating explainable AI tools and auditing for bias to ensure equitable predictions across diverse patient and payer groups.
- Explore multitask learning approaches to simultaneously predict procedure type and cost, enabling broader healthcare pricing insights.

References

- Health System Tracker (2024). How has U.S. spending on healthcare changed over time? Retrieved from https://www.healthsystemtracker.org/chart-collection/u-s-spending-healthcare-changed-time/
- Office of the Assistant Secretary for Planning and Evaluation (2021). Evidence on surprise billing: Protecting consumers with the No Surprises Act. U.S. Department of Health and Human Services. https://aspe.hhs.gov/sites/default/files/documents/no-surprises-act-brief.pdf
- Lopes et al. (2024). Americans' challenges with health care costs. KFF. https://www.kff.org/health-costs/issue-brief/americans-challenges-with-health-care-costs/
- Malik (2025). Enhancing healthcare cost transparency: Assessing implementation challenges and alternative solutions. Frontiers in Health Services, 4. https://doi.org/10.3389/frhs.2024.1379416