1. Title & Introduction

Title:

Predicting Total Hospital Charges for New York State Inpatient Discharges Using SPARCS Data and LightGBM Modeling

---

Introduction:

This project aims to predict total hospital charges for inpatient discharges in New York State using the SPARCS dataset. Accurate charge prediction can help hospitals manage healthcare costs, optimize resource allocation, and improve patient care planning. The modeling approach combines SQL-based data preparation, exploratory data analysis, feature engineering, and gradient boosting machine learning models — including segmented modeling and ensemble techniques — to achieve high predictive accuracy.

---

2. Dataset Overview

Dataset Source:

The data comes from the New York State Department of Health's SPARCS (Statewide Planning and Research Cooperative System) program.

Link to Dataset

Dataset Details:

- Approximately 2 million inpatient discharge records

- Features include patient demographics, diagnosis and procedure codes, length of stay, severity, admission type, payment type, hospital county, and total charges

- Due to size, the raw CSV is not included in the repo but can be downloaded from the official SPARCS site

Key Data Characteristics:

- Average total charge: ~$46,000 (range: $0.01 to ~$10.4 million)

- Length of stay grouped into short (0-1 days), moderate (2-3 days), long (4-7 days), and extended (8+ days)

- Charges vary substantially by age group, gender, diagnosis, procedure, admission type, and hospital county

---

3. Exploratory Data Analysis (EDA)

Key Trends and Insights on Hospital Charges and Patient Characteristics:

- **Charge Distribution:**

Total charges vary widely, with a median of about $24,889 and a mean of approximately $46,000, reflecting the skewness caused by high-cost outliers.

- **Demographic Patterns:**

Older patients (70+) incur the highest average charges (~$56,639), followed closely by the 50-69 age group. Male patients tend to have higher average charges than females ($50,751 vs $42,206).

- **Admission Types:**

Emergency admissions constitute the majority and have average charges ($46,700) slightly lower than elective admissions ($53,500).

- **Clinical Drivers:**

Common diagnoses include live births and septicemia. High-cost diagnoses like leukemias and heart valve disorders have significantly elevated charges, often exceeding $150,000.

- **Procedure Costs:**

Procedures such as organ transplants and tracheostomies are the most expensive, frequently costing over $400,000.

- **Hospital and Location Influence:**

Urban counties like Manhattan and Nassau have the highest average charges, while rural counties generally have lower costs.

- **Length of Stay Impact:**

Charges increase sharply with length of stay. Patients with stays longer than 8 days average over $120,000 in charges, compared to under $20,000 for 0-1 day stays.

---

4. Feature Engineering

Several Feature Engineering Steps Were Applied:

- **Categorical Encoding:** Patient demographics (gender, age group), clinical factors (severity, admission type, payment type), diagnosis and procedure descriptions, and hospital county were encoded numerically using label encoding or dense ranking to convert text categories into machine-readable formats.

- **Interaction Features:** Additional features were created by combining key variables (e.g., length of stay × severity, diagnosis × procedure) to capture complex relationships that can influence hospital charges.

- **Data Cleaning:** Missing or ambiguous values were handled by filtering or assigning default codes to ensure data quality.

---

5. Modeling Approach

- **Model Used:** LightGBM, a gradient boosting decision tree algorithm, was selected for its speed, efficiency, and strong performance on structured tabular data.

- **Segmentation:** The dataset was segmented by Length of Stay (LOS) into four groups — short, moderate, long, and extended — allowing tailored models to capture different cost dynamics for each segment.

- **Quantile Regression:** For long and extended LOS groups, quantile regression models were trained to better handle the skewness and heteroscedasticity in charges.

- **Hybrid Ensemble:** A hybrid ensemble combined predictions from LOS-specific models and a global model to achieve the best overall accuracy.

- **Evaluation Metric:** Mean Absolute Error (MAE) was used to evaluate model performance, reflecting the average absolute difference between predicted and actual charges.

---

6. Model Performance & Results

Model Variant	**MAE ($)**
Baseline Global LightGBM	~11,138
LOS Segmented Models (Short, Moderate, Long, Extended)	Varies by group (see below)
Quantile Regression (Long & Extended LOS groups)	Improved accuracy over baseline
Hybrid Ensemble (Quantile + Global)	~10,486

LOS Group MAE Breakdown:

LOS (length of stay) Group	MAE ($)
Short	7,274
Moderate	13,135
Long	23,671
Extended	28,091

The hybrid ensemble model blends specialized quantile regression models for long and extended LOS groups with the global model for other cases, achieving the best overall accuracy and reducing MAE by approximately 6% compared to the baseline.

Interpretation:

An MAE of around $10,486 indicates that on average, predicted charges deviate from actual charges by this amount. Considering the wide variability and complexity of inpatient costs, this represents a meaningful improvement and practical utility for cost forecasting.

---

7. Model Explainability & Insights

SHAP analysis was used to understand which features most influenced the hospital charge predictions:

- Length of stay and severity of illness were the strongest drivers of cost predictions across all models.

- Certain diagnosis and procedure combinations led to large spikes in predicted charges, especially in high-cost outliers (e.g., oncology-related

cases).

- Models trained on different length-of-stay groups showed distinct patterns in feature importance and prediction errors, supporting the use of segmented quantile regression models for long and extended stays.

- Additional features such as age group, gender, admission type, payment type, and hospital county had meaningful but smaller impacts on charges.

- Including interaction features between clinical and demographic variables improved model accuracy, indicating that combined effects are important for predicting costs.

---

8. Limitations

- **Data Limitations:** The dataset lacks some detailed clinical variables and insurance details that could improve prediction accuracy.

- **Model Limitations:** While LightGBM and segmentation improved performance, complex interactions and rare events may still cause prediction errors, especially in extreme cases.

- **Generalizability:** The model is trained on New York State data and may not directly generalize to other regions without adaptation.

- **Computation Time:** Some model training steps require long runtimes, which may affect scalability.

---

9. Future Work

- Incorporate additional features like patient comorbidities, insurance plans, and hospital-specific factors.

- Explore advanced ensemble models and deep learning architectures.
- Improve handling of rare diagnoses and outliers.
- Automate and optimize the data pipeline for faster retraining.
- Extend the model to predict other outcomes such as length of stay or readmission risk.

---

10. Conclusion

This project successfully leverages comprehensive healthcare data and advanced machine learning techniques to predict inpatient hospital charges with meaningful accuracy. The combination of segmentation, feature engineering, and ensemble modeling notably improved performance, highlighting the value of tailored approaches for heterogeneous patient groups.

While challenges remain—such as data limitations and computational demands—the insights gained provide a strong basis for future refinement and potential integration into healthcare cost management strategies.