Capstone Project

Hospital Billing Amount Prediction

Final Report

1. **Define the Problem Statement:** I set out to tackle the challenge of unpredictable hospital billing costs, which create inefficiencies in resource allocation—like overstaffing or underused beds—and financial stress for both hospitals and patients due to surprise bills. My goal was to use machine learning to predict billing amounts based on patient information, such as age, medical condition, and length of hospital stay. By accurately forecasting costs, I aimed to help hospitals plan resources more effectively, negotiate better terms with insurance providers, and potentially reduce financial overhead by 15-20%, aligning with industry benchmarks for cost-saving analytics. This solution addresses the critical need to streamline operations and improve affordability for patients.

2. **Model Outcomes or Predictions:** I designed this project as a regression task, meaning my models predict a specific dollar amount for each patient's hospital bill—a continuous number rather than a category. The expected output is a precise cost estimate based on patient details. I used supervised learning, where my models learn from historical patient records that include known billing amounts, to make these predictions accurately.

3. **Data Acquisition:** I used a single dataset, healthcare_dataset.csv, containing 55,500 patient records with details like age, medical condition, admission type, and billing amounts. I chose this dataset because it includes key factors directly tied to hospital costs, such as how long a patient stays or their diagnosis, making it ideal for my prediction task. While combining multiple data sources could provide broader insights, I found this dataset comprehensive enough for a focused project like mine. To ensure the data's suitability, I visualized its patterns early on. For instance, I created a box plot showing how billing amounts vary by medical condition—conditions like cancer often led to higher bills with more variability, confirming the data's ability to highlight cost drivers. I also used a correlation heatmap to see strong connections between features like length of stay and billing amount, which reassured me that the data could support accurate predictions.

4. **Data Preprocessing/Preparation:** I carefully cleaned and prepared the data to make it reliable and usable for my models, ensuring it was free of errors and formatted correctly.

a. **Techniques for Missing Values and Inconsistencies:** Although my dataset had no missing values overall, I added safeguards by filling any potential gaps in numeric fields like age or billing amount with the median (the middle value) to avoid skewing results, and in categorical fields like medical condition with the most common value. I removed 534 duplicate patient records (1% of the data) to avoid redundancy. I also found and removed negative billing amounts, which are invalid for hospital bills and likely data entry errors. To ensure consistency, I standardized text in the medical condition field by converting it to lowercase. I used box plots to spot unusually high or low billing amounts for each medical condition and removed outliers using the IQR method (keeping values within a reasonable range) to eliminate extreme cases that could confuse the models.

b. **Splitting Data into Training and Test Sets:** I split the data into an 80% training set, which I used to teach my models, and a 20% testing set to evaluate how well they predict on new, unseen data. I ensured this split was random but consistent using a fixed seed, making my results reproducible and fair.

c. **Analysis and Encoding Steps:** I created a new feature, "length of stay," by calculating the days between admission and discharge dates, and checked that all values were non-negative (removing any invalid ones). I scaled numeric features like age and room number to a standard range so models like Linear Regression or KNN wouldn't be biased by different scales. For categorical features like medical condition and admission type, I used one-hot encoding to turn categories into binary numbers (e.g., "cancer" becomes a 1 or 0 flag) that models can understand. I dropped irrelevant columns with too many unique values, like patient names and doctors, after confirming their high variety (e.g., 90% unique names) to prevent confusion. I also verified that age ranged from 13 to 89 years and billing amounts were positive after cleaning, ensuring data reliability.

5. **Modeling:** I selected nine supervised regression models to predict billing amounts, balancing simple and advanced approaches to find the most reliable solution. I started with Linear Regression as a simple baseline, assuming straightforward relationships between features like age and costs. I added Ridge and Lasso, which include penalties to prevent models from overcomplicating patterns—Ridge for stability and Lasso to focus on key features. I included Decision Tree to capture non-linear patterns, like sudden cost spikes for certain conditions. KNN predicts costs based on similar past patients, which is intuitive for hospital data. Random Forest and Bagging combine multiple decision trees to improve accuracy—Random Forest adds randomness to make trees diverse, while Bagging uses sampling for robustness. I also used Gradient Boosting and XGBoost, which build models step-by-step to correct errors, with XGBoost optimized for speed and performance. I chose these to compare basic versus sophisticated methods, ensuring I could identify the best predictor for hospital needs.

6. **Model Evaluation:** I focused on regression models to predict dollar amounts, as my goal was to estimate continuous billing costs using labeled data, rather than categorizing or finding patterns without labels. I evaluated models using three main metrics: Mean Squared Error (MSE, the average of squared prediction errors—lower is better for precision), Root Mean Squared Error (RMSE, errors in dollar terms—lower means smaller real-world mistakes), and R-squared (R^2, the percentage of cost variation explained—higher, closer to 1, shows better fit). For Random Forest and Bagging, I also checked Out-of-Bag (OOB) error, an internal measure of accuracy during training.

   To find the best model, I tested all nine on the 20% test set after optimizing their settings. Ensemble models—Random Forest, Gradient Boosting, and XGBoost—stood out with the lowest RMSE (errors in the thousands of dollars for a $2,000-$50,000 range) and highest R^2 (explaining most cost variation), showing they handle complex data patterns well. Linear models (Linear Regression, Ridge, Lasso) were simpler but less accurate, while Decision Tree and KNN performed moderately. I prioritized RMSE for practical dollar accuracy and R^2 for overall reliability, favoring interpretable models like Random Forest if scores were close. Visualizations supported this—scatter plots showed predictions aligning closely with actual bills for top models, residual plots showed no systematic errors, and error histograms were tightly centered around zero.

   a. **Interpretations:** My analysis of tree-based models (Random Forest, Gradient Boosting, XGBoost) revealed that Room Number, Age, and Length of Stay were the most influential factors driving billing amounts, far outweighing features like medical condition or medication. This means hospital costs depend heavily on how long patients stay and their basic demographics, not just their diagnosis or treatment.
   b. **Research Findings Summary:** I found that ensemble models like Random Forest and Bagging accurately predict billing amounts with errors in the thousands of dollars, making them reliable for hospital cost forecasting. The data was clean after removing duplicates and negative values, and visualizations confirmed that length of stay and age are key cost drivers, offering actionable insights for resource planning.
   c. **Next Steps Summary:** I plan to refine Random Forest and Bagging by further tuning their settings and possibly combining them with XGBoost for even better predictions. I'll deploy these models in a user-friendly app for real-time cost estimates, validate them on new data to ensure fairness, and test their impact in a hospital setting to achieve the targeted 15-20% cost reduction while prioritizing patient data privacy.