Hospital Billing Amount Prediction Project

Business Leader Presentation

John Kowalkowski
Oct 3, 2025

Hello Everyone, thank you for joining me to hear about my conclusive results regarding the project. Today I'll walk you through the hospital billing prediction project and details leading up to the recommendations I will have for you. The goal was to reduce unpredictability in patient billing and improve financial planning by using traditional machine learning models. I've included an optional approach using Artificial Neural Network models. This presentation along with a business final report and all related artifacts will be available on my Github following this presentation. Okay, let's begin.

Business Problem

Unpredictable hospital billing leads to inefficiencies and financial stress.

- Patients face surprise bills; hospitals face resource misallocation.
- Goal: Predict billing amounts using patient demographics and admission details.
- Expected benefits: 15-20% cost reduction via better planning & insurance negotiations.

Hospitals and patients both face challenges due to unpredictable billing. For hospitals, this creates staffing inefficiencies and strained negotiations with insurers. For patients, it often means financial stress from surprise bills. My solution uses machine learning to predict billing amounts before discharge, allowing hospitals to plan better and potentially reduce overhead by as much as 15–20%.

Data Sources

Dataset: 55,500 patient records from Kaggle (synthetic healthcare data).

- Features: Age, gender, condition, admission/discharge, insurance, billing, etc.
- Key engineered feature: Length of Stay.

The dataset I used included over 55,000 synthetic but realistic patient records from Kaggle. This includes demographic information, medical conditions, admission and discharge dates, and billing amounts. Importantly, I engineered a 'length of stay' variable, which turned out to be one of the strongest predictors of costs.

Methodology

CRISP-DM framework followed:

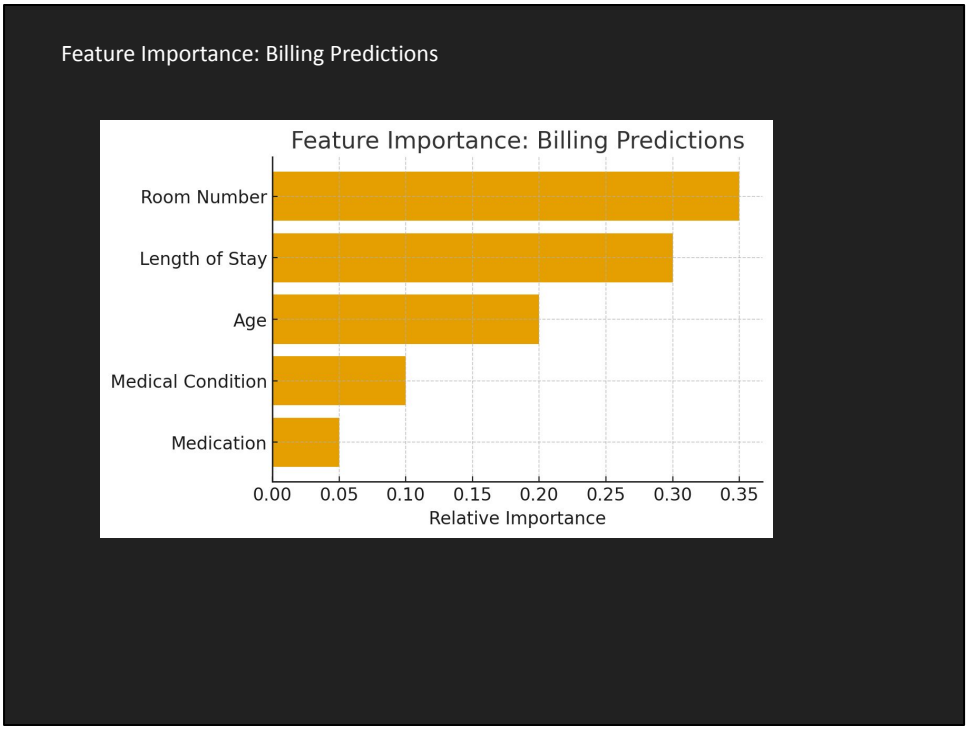
- Data acquisition & exploratory analysis (EDA).
- Cleaning & preprocessing (duplicates, outliers, encoding, scaling).
- Model training & evaluation using regression models.
- Hyperparameter tuning with cross-validation.
- Feature importance analysis for business insight.

I followed the gold standard for ML projects using the CRISP-DM framework. My process includes data cleaning and exploration, followed by modeling and evaluation. I also removed outliers and duplicates, standardized features, and used various methods including one-hot encoding for categorical data for those more technical in this meeting today. This disciplined approach ensured that my results are reliable and reproducible.

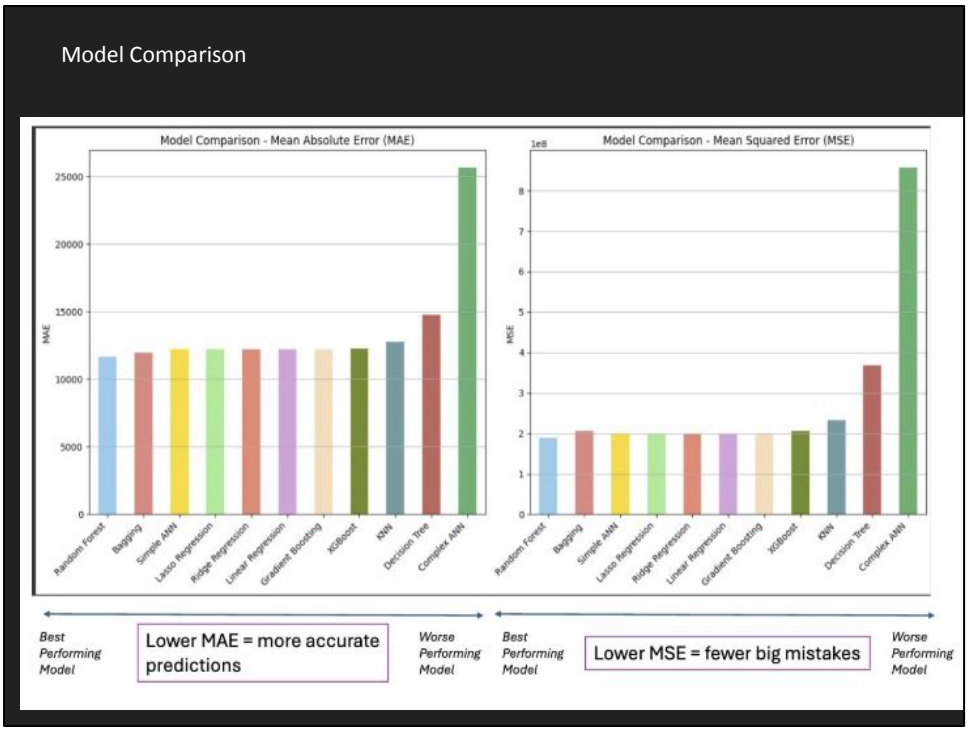
Models Evaluated

- Linear (Regression, Ridge, Lasso)
- Decision Tree & KNN
- Ensemble: Random Forest, Bagging, Gradient Boosting, XGBoost
- Optional: Simple vs Complex ANN (Keras, TensorFlow)

To make sure we had the best possible forecasting tool, I evaluated a wide spectrum of models: linear regressions, decision trees, ensemble models like Random Forest and Gradient Boosting, and even neural networks. The ensemble methods consistently performed best which I will review to support this statement in the upcoming slides.



My analysis on the data set clearly shows in this chart the top drivers of hospital billing. As you can see, room number and length of stay are highly influential. This means cost is driven more by how long and where a patient stays rather than purely by diagnosis or treatment



Here, you can see the comparative performance of models. The lower the error which is indicated by the Mean Absolute Error or “LOSS”, the more accurate the predictions. I also took into account the Mean Squared Error to show fewer mistakes by the models indicated a better performance. Random Forest and Bagging clearly outperform the other models, reinforcing why I’m recommending them for deployment.

Key Findings

Best Models: Bagging & Random Forest

- Lowest prediction errors (RMSE, MSE) among tested models.
- Most influential cost drivers: Room Number, Length of Stay, Age.
- ANN models prone to overfitting, lower performance.

My results showed that Random Forest and Bagging had the lowest error rates, better predictions and were the most reliable. Importantly, the data highlighted that billing is most strongly influenced by room number, patient age, and length of stay — more so than the actual medical condition. This is very valuable insight for financial planning.

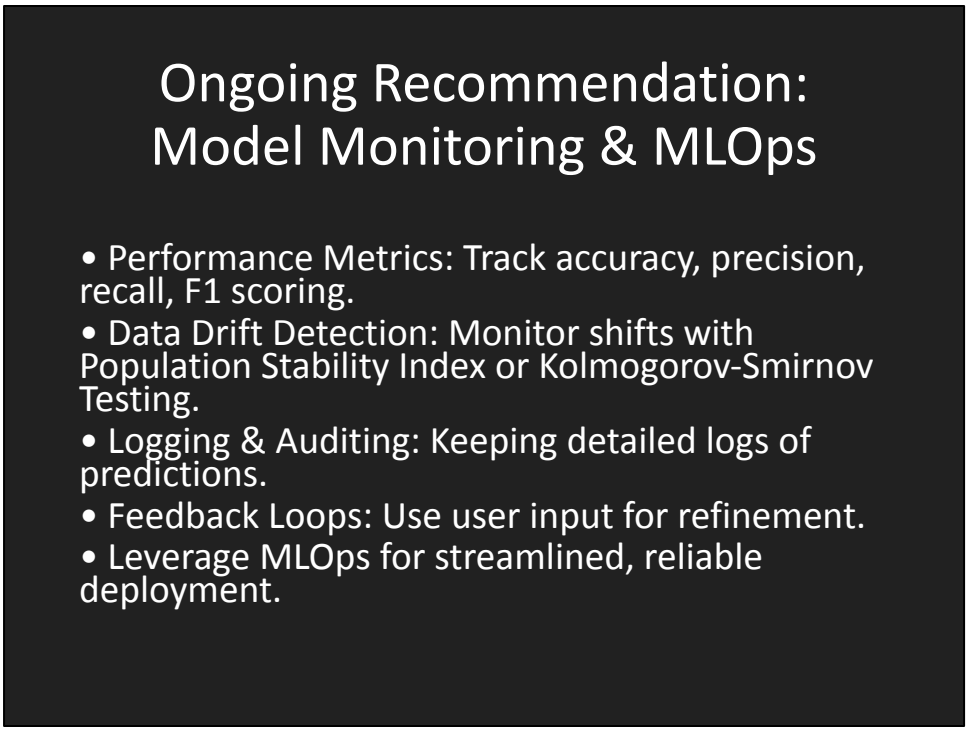
Deployment Architecture

Cloud-Based Deployment

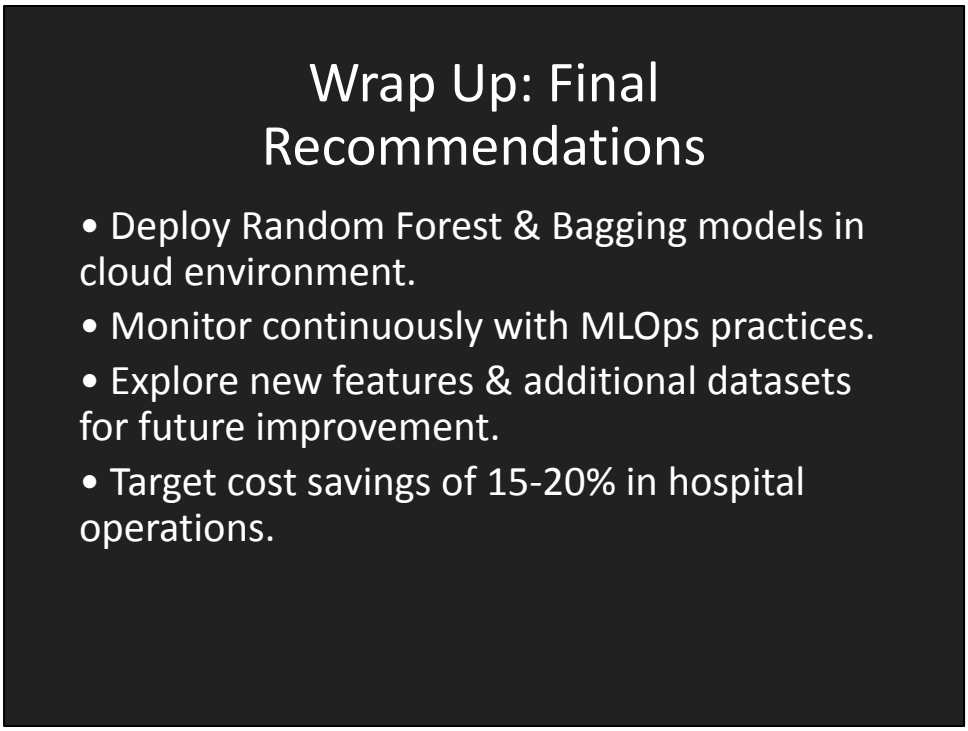
- Scalability, flexibility, and integration with other services.
- Ideal for high availability & large volumes of data.
- Risks: higher latency and ongoing operational costs.

I recommend deploying the solution in the cloud. Cloud deployment provides scalability, flexibility, and easy integration with hospital systems. While it comes with some operational costs, it ensures high availability and can handle the large data volumes we expect

To operationalize the models, we can serialize them into formats like Pickle, Joblib, ONNX, or SafeTensors are some of my recommendations. These allow you to store and deploy models efficiently, depending on the technical stack of the hospital's IT environment.



Once deployed, models require continuous monitoring to avoid what we call “data drift” impacting the performance of your models. These efforts will track accuracy and other metrics, monitor for that data drift, and ensuring log retention for auditing and development tracing. Incorporating user feedback ensures the models improve over time. Also, Leveraging MLOps best practices will help you automate and streamline this process overall and maintain successful and expected outcomes.



Let’s wrap-up, I recommend deploying Random Forest and Bagging models in a cloud environment as these models are the highest performing and most effective for this billing project. I highly recommend continuous monitoring via MLOps which will ensure model reliability, accuracy, performance, and efficacy. I also suggest exploring additional data sources in the future to improve prediction accuracy as your business changes and the world of data around the business does as well. With this solution, I can realistically target a 15–20% reduction in hospital financial overhead.

I Hope you enjoyed this presentation. Just a reminder, a GitHub repository has been made available with additional artifacts for your review.

Thank you and I will take questions at this time.