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.

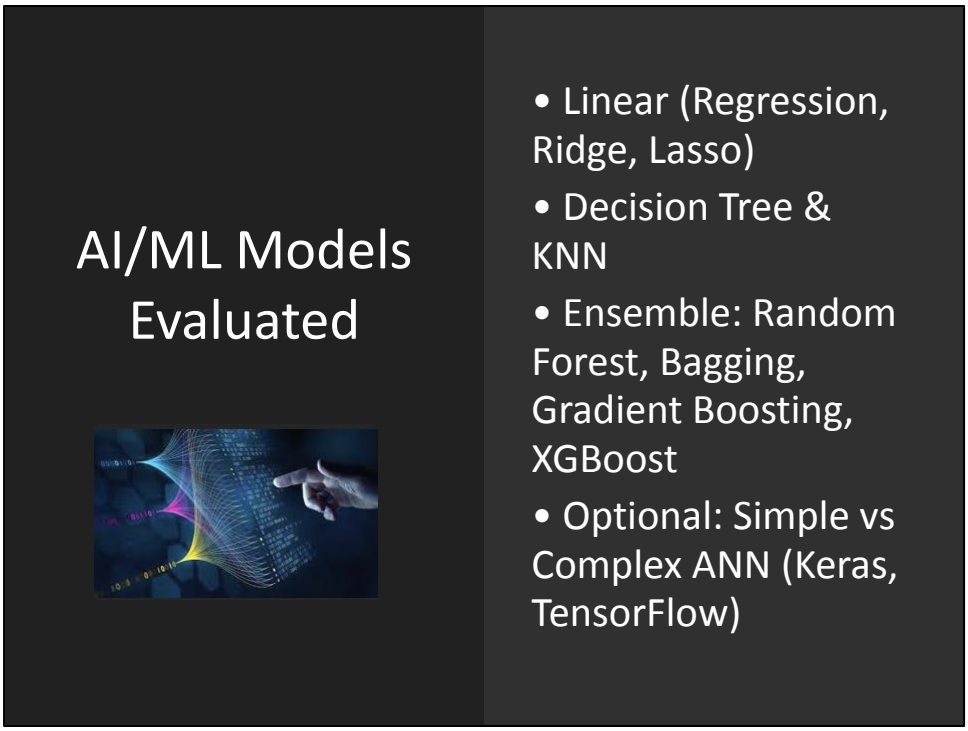
The dataset I used included over 55,000 synthetic but realistic patient records from Kaggle. This includes demographic information, medical conditions, admission and discharge dates, and billing amounts. Importantly, I engineered a 'length of stay' variable, which turned out to be one of the strongest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

Methodology

"CRISP-DM Framework"

- 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.

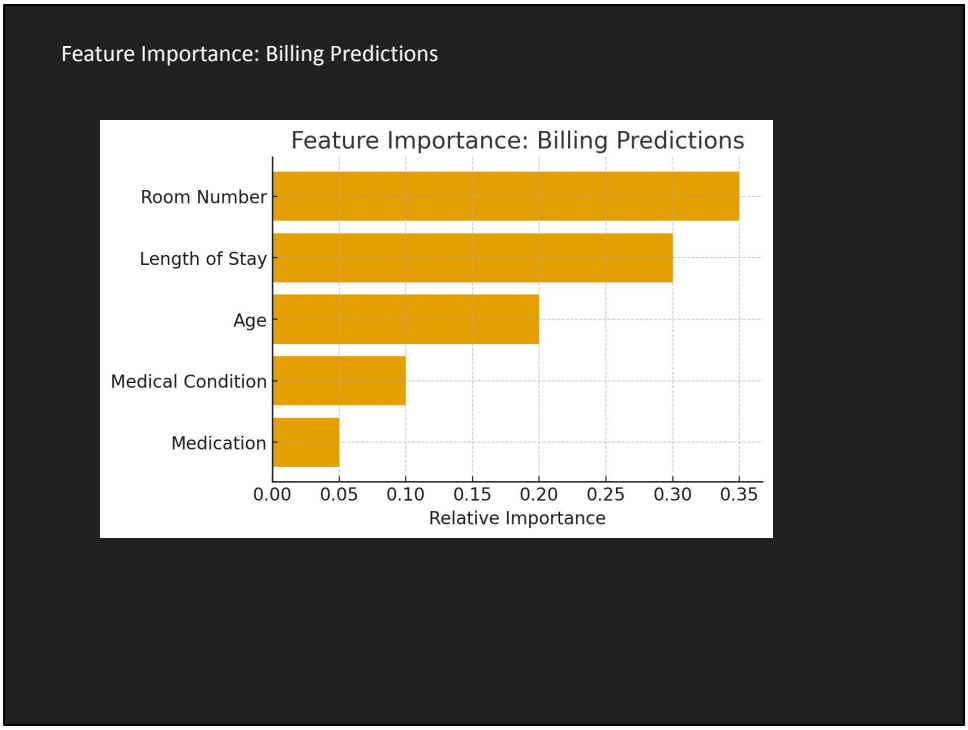


To make sure we had the best possible forecasting tool, I evaluated a wide spectrum of models: linear regressions, decision trees, ensemble models like Random Forest and Gradient Boosting, and even neural networks.

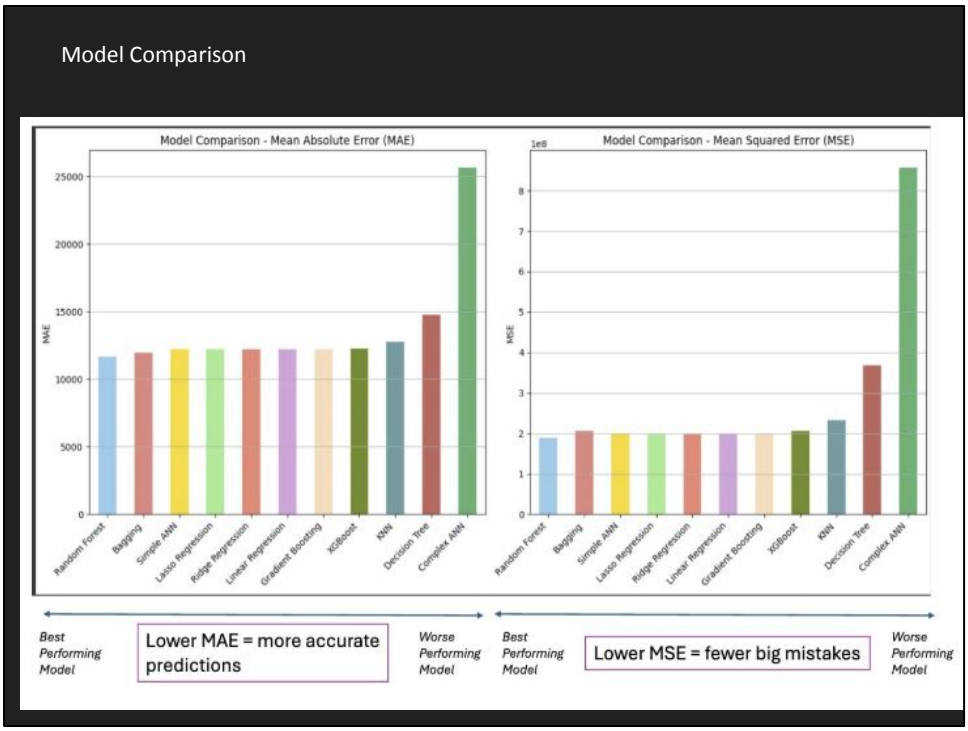
To identify the best forecasting tool, we tested several AI models — each offering different strengths. Let me share with you at a high-level what these mean as to give you a better understanding conceptually.

- **Linear Regression:** A simple, transparent model that helps us see basic relationships in our data, like how billing patterns change with patient volume.
- **Decision Trees:** These work like a series of business rules — easy to interpret and explain how predictions are made.
- **Random Forest:** Combines many decision trees to get more stable, accurate forecasts — reducing noise and improving reliability.
- **Gradient Boosting:** Builds trees step by step, learning from past errors to fine-tune results — great for complex forecasting.
- **Neural Networks:** Modeled after the human brain, these can capture very complex patterns but act more like a “black box.”

Across all, the ensemble models — Random Forest and Gradient Boosting — delivered the strongest, most consistent results, and I'll show their performance in the next slides.



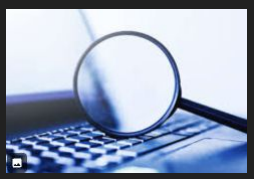
My analysis on the data set clearly shows in this chart the top drivers of hospital billing. As you can see, room number and length of stay are highly influential. This means cost is driven more by how long and where a patient stays rather than purely by diagnosis or treatment



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.

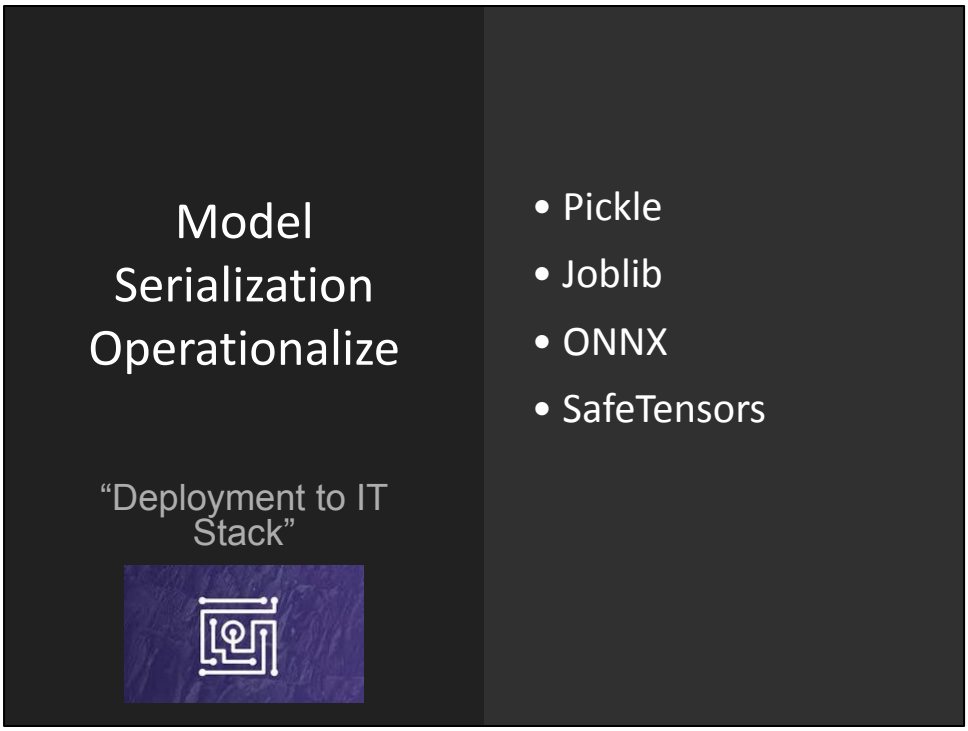
My results showed that Random Forest and Bagging had the lowest error rates, better predictions and were the most reliable. Importantly, the data highlighted that billing is most strongly influenced by room number, patient age, and length of stay — more so than the actual medical condition. This is very 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 move these models from development into daily use, we need to operationalize them — in other words, make them ready for real-world deployment.

I can package or "store" the models in standardized formats like Pickle, Joblib, ONNX, or SafeTensors that make them easy to integrate into the hospital's existing IT systems.
This ensures they run efficiently, stay secure, and can be updated as new data becomes available.

The specific format we use will depend on the hospital's technical environment, but the goal is the same — reliable, scalable deployment that fits seamlessly into operations.

Model Monitoring & MLOps

"Continuous Monitoring"

- Performance Metrics: Track accuracy, precision, recall, F1 scoring.
- Data Drift Detection: Monitor shifts with Population Stability Index or Kolmogorov-Smirnov Testing.
- Logging & Auditing: Keeping detailed logs of predictions.
- Feedback Loops: Use user input for refinement.
- Leverage MLOps for streamlined, reliable deployment.

Once the models are deployed, the work doesn't stop — they need continuous monitoring to ensure accuracy and reliability over time.

You would need to track key performance metrics to confirm the models are predicting correctly and meeting expected outcomes.
I would also recommend monitoring for data drift, which happens when real-world data changes — for example, new patient behaviors or billing patterns — so the model can adapt before accuracy declines.

It's a crucial to maintaining detailed logs allowing for full transparency and auditing, while user feedback loops help improve the model learning capabilities and adjust continuously.

Finally, by applying MLOps best practices, you can automate and streamline these processes — ensuring the system remains stable, compliant, and delivers consistent business value.

Final Thoughts

"Recommendations and Next Steps"

- Deploy Random Forest & Bagging models in cloud environment.
- Monitor continuously with MLOps practices.
- Explore new features & additional datasets for future improvement.
- Target cost savings of 15-20% in hospital operations.

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.