Healthcare Utilization Prediction Using Machine Learning

Predicting Patient Hospitalization Risk from Claims Data

Mukund Valluru

M.S. Business Analytics
Rochester Institute of Technology

Project Summary

This project applies machine learning techniques to healthcare claims data to predict patient utilization, measured as expected days of hospitalization. The work focuses on feature engineering, model comparison, and evaluation to support data-driven healthcare risk assessment and resource planning.

Tech Stack
Python • Pandas • Scikit-Learn • Neural Networks • Statistical Modeling

1 Introduction

The Heritage Health Prize dataset provides claims data for predicting a patient's number of days in the hospital. Our aim is to clean, transform, and engineer relevant features before employing different machine learning models—including Linear Regression, Random Forest, and a Simple Neural Network with Ridge regularization—to assess model performance.
This report covers:

- Which variables were dropped (and why).
- How missing data were handled.
- The transformations performed on certain variables.
- Modeling and results, including performance metrics and key plots.

2 Dropped Variables

Based on exploratory analysis and domain considerations, we removed:

- **MemberID:** Unique identifier, not predictive.
- **ProviderID:** Overly granular; minimal impact on target.
- **Vendor:** Billing company does not have predictive value.
- **PCP:** Specific doctor IDs does not improve predictions.
- **SupLOS:** Became unnecessary after careful handling of missing LengthOfStay data.

Rationale: Excluding variables that add noise or irrelevance can help models focus on more predictive signals.

3 Handling Missing Values

To preserve as much information as possible, we used the following strategies:

- **AgeAtFirstClaim:** Imputed missing values with median age.
- **Sex:** Converted nulls to "Unknown" category.

- **PayDelay:** Imputed missing with median PayDelay; capped extreme outliers at 365 days.
- **LengthOfStay:** If truly missing and not flagged by SupLOS, assumed 0; if flagged, indicated suppressed data but retained the record.

Importance: These choices ensure the data remain as complete as possible while preventing spurious or biased signals.

4 Data Transformations

Several transformations were used to handle categorical and numeric variables:

- **Binning Skewed Variables:** Grouped PayDelay into buckets (e.g., 0–10, 10–20, etc.) to reduce the impact of extreme values.
- **Encoding Categorical Variables:** Created dummy variables for Sex, Specialty, PrimaryConditionGroup, etc.
- **Grouping Rare Categories:** Combined uncommon specialties and procedure groups into an "Other" category.

Why: These steps yield cleaner, more model-friendly data, reduce noise, and handle skewness effectively.

5 Predictive Modeling and Results

We evaluated three models on the transformed dataset:

5.1 Linear Regression

- **Method:** Ordinary Least Squares (OLS) regression
- **Metrics:**
  - *RMSE (Root Mean Square Error):* 1.52
  - *R^2:* 0.0635

A sample plot comparing predicted vs. actual days in the hospital is shown in Figure 1.

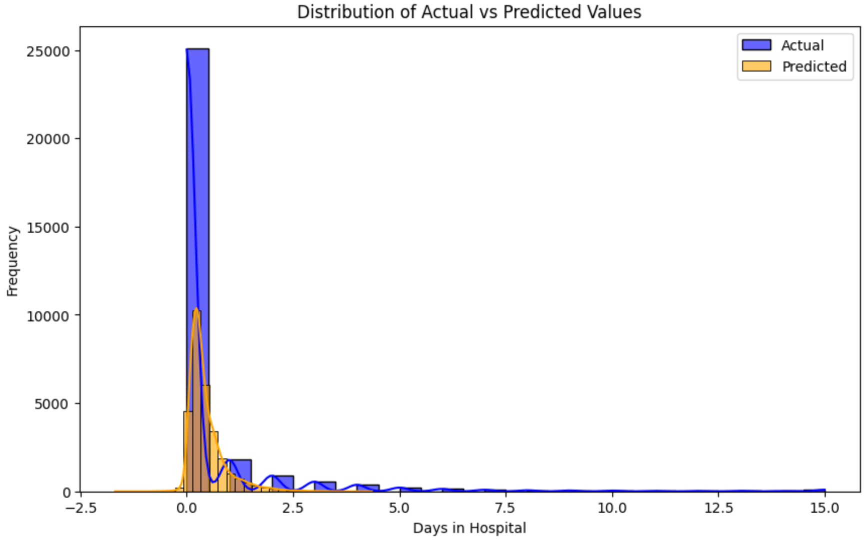

Figure 1: Linear Regression Predicted vs. Actual Days in Hospital

5.2 Random Forest

- **Metrics:**
  - *RMSE:* 1.55
  - *R^2:* 0.0203

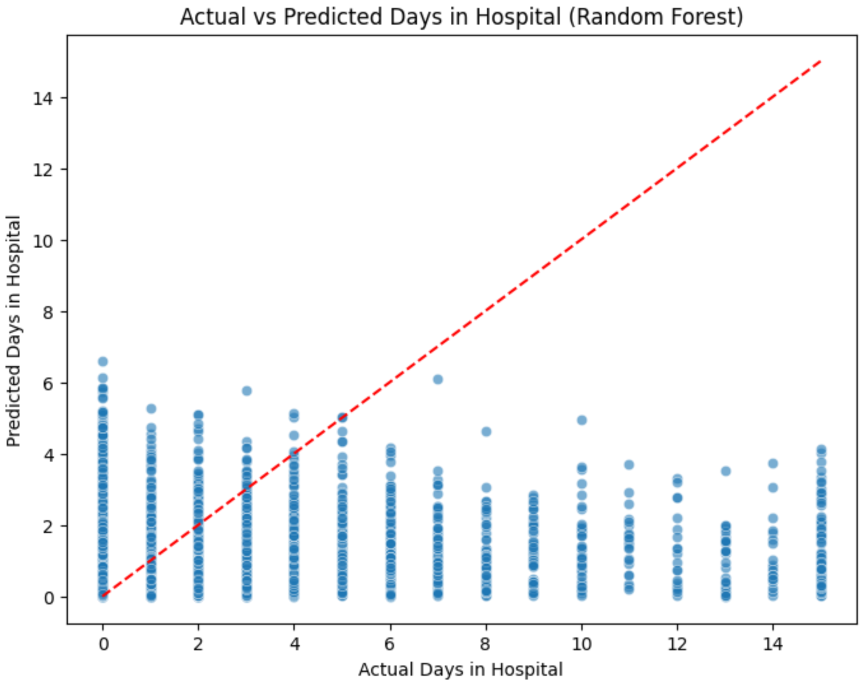

Figure 2: Random Forest Actual vs Predicted Values

Figure 2 illustrates how the model performed by comparing the actual values vs the predicted values.

5.3 Simple Neural Network with Ridge Regularization

- **Architecture:**
  - three hidden layers - 128,64,32 neurons
  - Activation function: ReLU
- **Ridge Regularization (ℓ_2):** Applied to the weights to prevent overfitting.
- **Metrics:**
  - *RMSE:* 1.51
  - R^2*:* 0.0689

Below, Figure 3 shows the actual values vs predicted values.

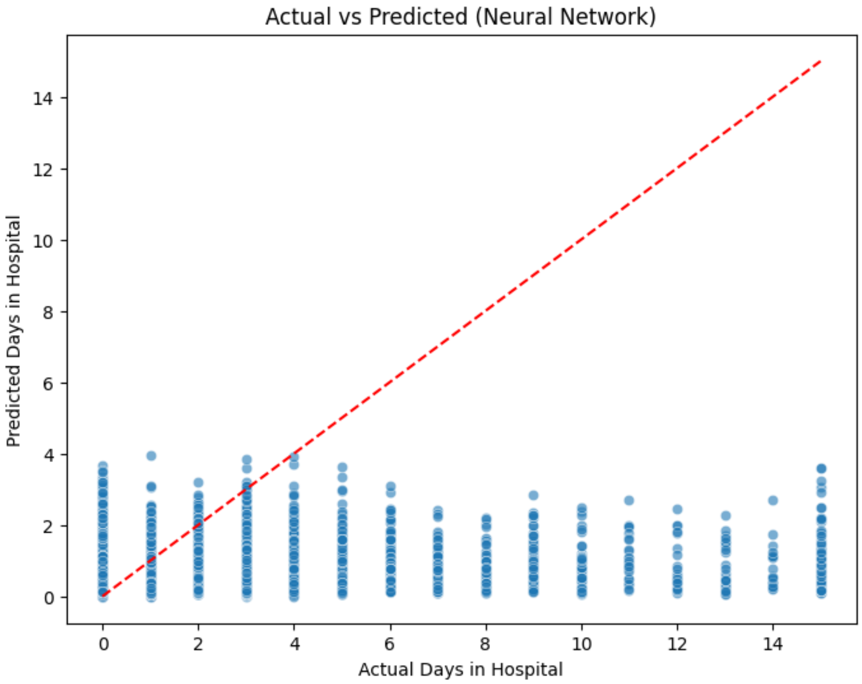

Figure 3: Neural Network Actual vs Predicted Values

Observations: The simple neural network slightly outperforms Linear Regression and Random Forest in terms of RMSE and R^2, suggesting that the model can capture non-linear patterns in the data—provided we have enough data and proper regularization.

5.4 Summary of Model Performance

Model	**RMSE**	R^2
Linear Regression	1.52	0.0635
Random Forest	1.55	0.0203
Neural Network (Ridge)	1.51	0.0689

Table 1: Comparison of Model Results

6 Conclusion

After thorough data preparation—dropping irrelevant columns, imputing missing values and transforming skewed variables—we tested three prediction models (Linear Regression, Random Forest, and a Simple Neural Network). The Neural Network with ridge regularization performed better.