Rancel Hernandez

D214 Data Analytics Graduate Capstone

April 30, 2025

Executive Summary

The research question for this analysis was: Can a random forest regression model accurately predict patient admission duration based on demographic data, clinical measurements, and admission information from a publicly available critical care database? This question was chosen to evaluate whether random forests can improve hospital admission estimates compared to current methods. A more accurate model could allow rooms to be prepared earlier, staff to be scheduled more efficiently, and patients to receive better estimates of their stay, significantly improving hospital resource allocation and patient outcomes. The null hypothesis is that the model cannot predict admission duration within a root mean squared error (RMSE) of 5 hours. This ambitious threshold was selected to prioritize precision and aim for an ideal goal due to having limited clinical context prior to data preparation.

This analysis uses a publicly available critical care dataset from a hospital in Boston, Massachusetts, containing de-identified patient information. The focus is specifically on admissions from the emergency department because these are typically unplanned and contribute significantly to hospital congestion. The dataset includes features such as diagnostics, vital signs, procedures, medication information, and admission details, all relevant to modeling admission duration.

However, due to excessive missing values and other data quality issues, all the features could not be used together in a single dataset. Therefore, multiple datasets were derived from the original source, each containing a different subset of variables. As shown in Table 1, this includes a base dataset with 544,190 observations and three specialized subsets with more detailed clinical features. This approach enabled broader use of the available data while allowing for targeted exploration of specific factors contributing to admission duration.

Derived Subset	Observations	Features	Exclusive Features
Base	544,190	26	N/A
Vitals	89,669	30	max_bmi, min_bmi, weight, and height
Blood Pressure	48,928	29	systolic blood pressure (SBP), diastolic blood pressure (DBP), and mean arterial pressure (MAP)
Procedures	162,992	27	procedure_category

Table 1

The data preparation process began by downloading the relevant files from the dataset's publisher and loading them into corresponding tables in pgAdmin, with appropriate column names and data types. An exploratory analysis was conducted to identify usable features, and several new features were engineered, including medication delays, procedure counts, readmission status, and MAP. Features with significant missing values were excluded rather than imputed to maintain data quality.

To reduce dimensionality, categorical features with excessive unique values were aggregated. For example, the language feature was reduced from 25 categories to the 5 most

common languages, with less frequent ones grouped into an "Other" category. One-Hot Encoding was then applied to prepare categorical features for modeling, while ordinal categorical variables, such as drg_severity, were already numerically encoded.

Outliers were removed because clinical data often contain severe data entry errors, such as a medication duration of 950 years. These values were excluded to prevent certain trees in the ensemble from skewing the final predictions. Outliers were detected using Z-scores and the standard three-standard-deviation threshold, which is most appropriate for features that follow a normal distribution. Although not all features were normally distributed, this approach was still applied because values beyond three standard deviations were rare and clearly deviated from the majority of the data. Domain-specific thresholds were also applied to features with physiological limitations. For example, body mass index (BMI) values were restricted to a range of 10 - 100 kg/m^2. Additionally, some records were removed due to impossible values, such as negative admission durations.

Each dataset was split into training and test sets, with a consistent test ratio of 0.2 across all datasets. The base dataset used a reduced training ratio of 0.3 to shorten training time because additional observations did not improve accuracy but significantly increased computational cost. The remaining subsets used a standard 0.8 training size. A separate validation set was not required because out-of-bag (OOB) scores from the Random Forest models served as an internal validation metric.

A Random Forest Regressor was chosen to model admission duration because it can capture complex and nonlinear relationships between features and outcomes, which is crucial given the high variability in clinical data. This includes handling interactions between categorical

and continuous features. Additionally, Random Forest makes minimal assumptions about the data, offering greater flexibility compared to more rigid models.

Random Forest models rely on three main assumptions: limiting irrelevant features, having sufficient data, and predicting within the range of the training data. Too many irrelevant features can introduce noise and slow training, but this risk was minimized through thorough feature selection. A large dataset is also required because each tree is trained on a subset of the data, which was not an issue given that the base dataset had over 500,000 records and the smallest subset still had around 50,000. Like other tree-based models, Random Forests can't extrapolate well beyond the training range, but this wasn't a concern because admission durations were expected to fall within the observed values.

GridSearchCV was used to tune the Random Forest hyperparameters based on three-fold cross-validation with negative MSE scoring. The parameters tuned included tree depth, minimum samples per split and leaf, number of estimators, and maximum features. The values tested were chosen to reduce overfitting while preserving model flexibility. The best-performing combination was 300 estimators, a max depth of 20, a minimum samples split of 10, and a minimum samples leaf of 1. Although this configuration slightly improved performance, the difference in RMSE compared to the default model was only one hour, so further tuning was not done as it was unlikely to improve performance.

The Random Forest models significantly outperformed the baseline model, which predicted the test set's average admission duration and had an RMSE of approximately 89 hours. The best-performing model, the vitals model, achieved a test RMSE of approximately 50 hours, representing a 44% improvement over the baseline. The base model followed closely with a test

RMSE of 52 - 53 hours, while the blood pressure and procedures models performed worse, at 63 and 71 - 73 hours, respectively. All the RMSE scores are shown in Table 2.

Model RMSE Scores	Initial Train Set	Initial Test Set	Absolute Difference	Reduced Train Set	Reduced Test Set	Absolute Difference
Base	30.06	52.23	22.17	41.72	53.1	11.38
Vitals	21.60	49.55	27.95	33.64	50.9	17.26
Blood Pressure	27.65	63.09	35.44	44.88	63.69	18.81
Procedures	39.62	71.26	31.64	53.58	72.75	19.17

Table 2

Despite this improvement, none of the models met the hypothesis threshold of 5-hour RMSE, so we fail to reject the null hypothesis that the random forest model cannot predict admission duration within this margin. The goal was overly ambitious given the target variable's high standard deviation of approximately 100 hours.

The vitals model also achieved the highest test R-squared score at approximately 0.72, meaning it explained 72% of the variance in admission duration. This indicates strong performance considering the inherent variability and complexity of modeling hospital stays. The other models had R-squared values between 0.48 and 0.7, with the procedures model performing the worst. All the R-squared scores are shown in Table 3.

Model R-squared Scores	Initial Train Set	Initial Test Set	Absolute Difference	Reduced Train Set	Reduced Test Set	Absolute Difference
Base	0.89	0.65	0.24	0.78	0.64	0.14
Vitals	0.95	0.72	0.23	0.87	0.71	0.16
Blood Pressure	0.94	0.66	0.28	0.84	0.66	0.18
Procedures	0.83	0.49	0.34	0.70	0.47	0.23

Table 3

The feature importance analysis showed that medications_ordered was the most influential feature in the best-performing models, contributing 42 - 66% of the total importance. Other key features included drg_severity, diagnoses_count, age, and procedure_count, though each accounted for roughly half or less of that importance. In contrast, the procedures model, which performed the worst, ranked medications_ordered seventh, suggesting that models assigning greater weight to this feature consistently achieved better performance. The feature importances of the initial set are visualized in *Figure 1*.

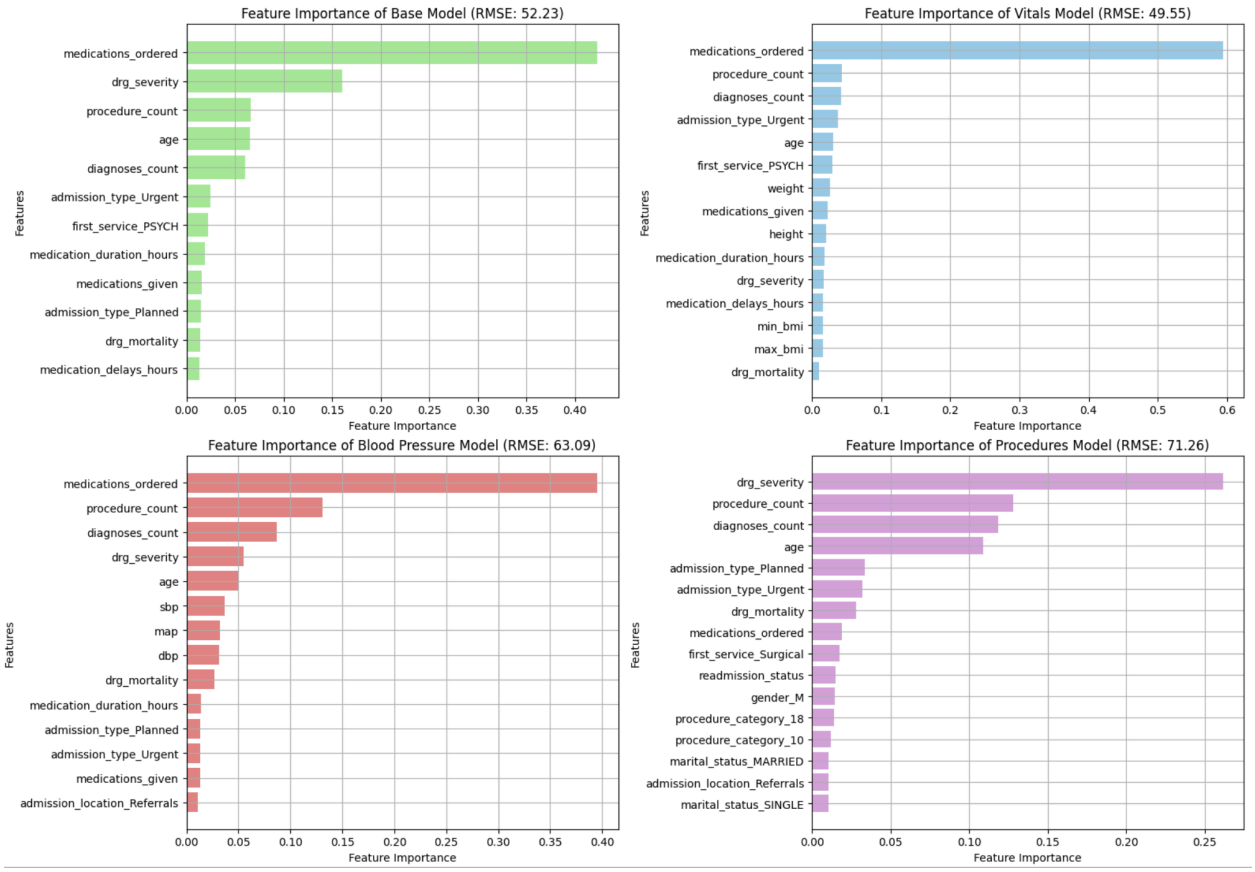

Figure 1: Comparison of feature importances across all evaluated random forest models.

The initial models showed signs of overfitting, with training-test RMSE gaps exceeding 20 hours. To address this, a feature importance threshold of 1% was applied, which removed roughly 75% of the features per model. After this reduction, the gap between train and test scores decreased significantly without harming test performance, as shown in Tables 2 and 3. This confirmed that many of the removed features were introducing noise rather than contributing predictive value. Additionally, the top feature importances also remained consistent between the initial and reduced sets, with the importance of removed features redistributed to the stronger predictors. The OOB and test set scores for RMSE and R-squared were consistently similar, with differences under 2%. This consistency reinforces that the training and test sets were properly

randomized and followed similar distributions. The comparison of OOB scores is shown in Table 4.

Model	Initial RMSE OOB Set	Reduced RMSE OOB Set	Initial R-squared OOB Set	Reduced R-squared OOB Set
Base	52.39	53.14	0.654	0.64
Vitals	49.69	50.83	0.73	0.71
Blood Pressure	65.13	65.94	0.66	0.65
Procedures	69.75	70.84	0.48	0.47

Table 4

The analysis faced several limitations related to model behavior, data structure, and the tools and techniques used. One major challenge was overfitting in the initial feature sets, with training-test RMSE gaps exceeding 20 hours. The reduced feature set removed many of the weaker predictors and significantly narrowed this gap without affecting test accuracy. However, some overfitting remained, indicating that further improvement may require a more targeted feature selection or regularization techniques.

The original data source included categorical features with many unique values, which led to high dimensionality after One-Hot Encoding. To address this, category aggregation was used to reduce the number of generated binary columns. While this significantly reduced dimensionality, it may have obscured meaningful distinctions between grouped categories, particularly in features like race and language.

Data preparation also introduced limitations, especially when working with large tables. For example, querying the vital signs table in pgAdmin, which contained over seven million

entries, resulted in noticeably slower performance. Additionally, linking vital signs and other clinical data required careful join logic and filtering, adding complexity and time to the preprocessing stage.

The tools and programming libraries used in this analysis also introduced technical limitations. PostgreSQL provided clean syntax and efficient data handling but occasionally selected suboptimal join strategies, which increased runtime for complex queries. The core analysis was conducted in Python, which offered flexibility and access to essential libraries like Scikit-Learn. However, Python was not ideal for large-scale data manipulation, so most operations were handled more efficiently in PostgreSQL. Pandas was necessary for data analysis but had less intuitive syntax for certain tasks, often requiring additional time spent reviewing documentation.

GridSearchCV was used to optimize the Random Forest models, but the process was computationally intensive. Each forest contains hundreds of trees and is retrained across multiple folds and parameter combinations, making the grid search slow. In this analysis, tuning took several minutes per iteration, which made it inefficient for experimenting with multiple configurations.

The techniques used in the evaluation process also had limitations. The feature importance analysis identified which features were most impactful but did not explain the direction or nature of their influence. For example, knowing that patient age is important does not clarify whether older patients tend to stay longer or shorter. Similarly, RMSE was useful for measuring the average error but did not indicate whether the model tended to overpredict or underpredict. This distinction is critical in healthcare because overestimates could lead to resource inefficiencies, while underestimates could disrupt discharge planning.

Finally, the analysis was based on data from a single hospital, which limits how well the findings can be generalized to other healthcare settings. Differences in patient populations, care protocols, and operational practices across hospital networks were not captured in the dataset. While the vitals model achieved strong performance with an R-squared of 0.72, the remaining 28% of unexplained variance likely reflects external factors not included in the available features. Given the high variability in admission duration, even the best model's RMSE of 50 hours, which is approximately 2.1 days, is a substantial margin of error. As a result, the model's accuracy would likely decrease when applied to patients in different hospitals unless it is retrained or validated on data from additional institutions.

Based on the results of the analysis, three key actions are recommended to improve admission duration modeling and support future research: optimizing the database structure, analyzing medication regimens in greater depth, and integrating multi-institution data. First, the hospital should invest in improving data gathering and database structure. Although the dataset was comprehensive in scope, it was not designed for admission prediction. Over 60% of the analysis time was spent on data extraction and preparation, with many features requiring extensive engineering or being excluded due to missing or messy values. For example, vital signs were not directly tied to specific admissions and had to be manually linked using patient IDs and timestamps. These entries were also mixed with outpatient data, increasing the complexity of preprocessing. Optimizing the structure of these features and standardizing data formats would reduce preparation time and improve the overall quality of the inputs available for modeling.

This recommendation is realistic because the necessary data already exists within the system and simply requires better integration and management. Steps could include improving how different clinical systems exchange information, hiring database engineers to maintain the

architecture, and adopting standardized formats to ensure consistency. These efforts would support not only this model but also future analyses that rely on clinical data, reducing preparation time and associated costs.

Second, future research should investigate how medication regimens influence admission duration. Medications_ordered was consistently one of the most important predictors in the best-performing models. Specific medications were not included in this analysis due to the high dimensionality of the feature. Aggregating medication types or performing a market basket analysis could uncover patterns in common drug combinations and their relationships with length of stay. This could lead to insights into how certain regimens affect recovery time, improve admission duration modeling, and help guide prescribing practices.

Third, integrating data from other institutions would help build a more generalizable model. The current analysis was limited to a single hospital, which restricts its usefulness across diverse healthcare settings. Combining this dataset with records from additional hospitals would support validation across institutions and help identify universal predictors of admission duration.

Implementing the proposed actions would offer several concrete benefits, both operational and clinical. The most immediate improvement is a substantial reduction in uncertainty when estimating admission duration. The baseline model, which predicted the average duration, had a RMSE of approximately 89 hours, or 3.8 days. In contrast, the vitals model achieved an RMSE of 50 hours, reducing the error margin by 44%, which represents a difference of nearly 1.7 days. This reduction in uncertainty could allow clinicians to provide more accurate discharge estimates, improving communication with patients and families and potentially increasing patient satisfaction.

More accurate predictions also support better resource planning, including improved room assignments, timely room preparation for incoming patients, more efficient staff scheduling, and reduced emergency department overcrowding caused by delayed discharges. These improvements could directly enhance operational efficiency, especially during high-volume periods when staff are overwhelmed. Additionally, future medication regimen analysis could lead to insights that improve prescribing practices. If certain medications or combinations are shown to shorten hospital stays, those patterns could inform treatment protocols and contribute to shorter admissions, improving patient outcomes and reducing strain on hospital resources.

Finally, collaboration between institutions through multi-hospital data integration would create opportunities for knowledge sharing and the standardization of best practices in patient care management. Beyond improving model accuracy, this approach could establish a framework for healthcare systems to collectively address shared challenges in resource optimization and care delivery. These recommendations have the potential to strengthen hospital operations, improve patient outcomes, and generate insights that benefit healthcare institutions nationwide.

In conclusion, this analysis demonstrates that Random Forest regression models can significantly improve admission duration predictions compared to baseline methods, achieving a 44 percent reduction in error. Although the ambitious 5-hour threshold was not met, the findings identified key predictors of hospital stays and established a foundation for future modeling efforts. By implementing the recommended database optimizations, expanding medication analysis, and fostering cross-institutional collaboration, hospitals can leverage these insights to improve resource allocation, increase operational efficiency, and enhance the overall quality of patient care.