**Predicting Hospital Length of Stay**

**Data Exploration Findings**

1. **Gender and Age Group Trends**: I discovered that males tend to have a longer length of stay (LOS) across all age groups. Interestingly, patients aged 50-69 actually have a longer average LOS compared to those aged 70 and older, which was a bit surprising.
2. **Admission Type Impact**: The analysis showed that elective admissions result in longer stays than emergency or trauma admissions. It’s also noteworthy that when the emergency department indicator is "No," the average LOS tends to be longer.
3. **Procedure Impact**: Tracheostomy procedures stand out significantly, with patients having much longer LOS—averaging 55 and 44 days—compared to all other procedures.

**Predictor Variables and Dataset Dimensions:**

I cleaned the data by replacing blanks and null values with "Unknown," and I removed certain rows that had too few values for a proper 70:30 train-test split. This included exclusions for specific gender, payment typology, and APR DRG codes. The final dataset now has **28,095 observations**, with a comprehensive list of potential predictor variables like:

* Hospital Service Area
* Age Group
* Gender
* Race
* Ethnicity
* Type of Admission
* APR DRG Code
* APR Severity of Illness Code
* APR Risk of Mortality
* APR Medical Surgical Description
* Payment Typologies 1-3
* Emergency Department Indicator
* Total Charges
* Total Costs
* Patient Disposition

I created two lists of potential predictors: one that includes Total Charges, Total Costs, and Patient Disposition (df\_notreduced) and another that excludes them (df\_reduced), since I expect that information to be unavailable when a patient is first admitted.

**Model Performance:**

I conducted a 70-30 train-test split and compared the performance of different models:

**Models with df\_reduced:**

|  |  |  |
| --- | --- | --- |
| **Model** | **Testset RMSE** | **Testset R²** |
| Linear Regression | 9.493 | 0.243 |
| Decision Tree | 9.369 | 0.262 |
| Random Forest | 9.359 | 0.264 |

**Models with df\_notreduced:**

|  |  |  |
| --- | --- | --- |
| **Model** | **Testset RMSE** | **Testset R²** |
| Linear Regression | 6.514 | 0.643 |
| Decision Tree | 5.958 | 0.701 |
| Random Forest | 5.654 | 0.731 |

**Model Insights:**

For the models using df\_reduced, I found that the RMSE was around 9.5 days, which means there’s quite a bit of inaccuracy in predicting the duration of stay, and they only explained about 25% of the variability—definitely not satisfactory.

On the flip side, the models using df\_notreduced showed much better performance, with an average RMSE of around 6 days and an R² of about 70%. However, I noticed that the top two features driving this model were Total Costs and Total Charges, which is a bit concerning. It suggests that without this information, the model wouldn’t be very useful for initial predictions of LOS, going against my goal of using reliable surrogates.

**Suggested Improvements:**

To enhance the model's performance, I believe we need to focus on improving data collection first. Gathering more detailed patient health data at admission—like heart rate, blood pressure, and mobility—could really help improve accuracy.

Next, it would be beneficial to develop specialized models for different APR DRG descriptions. The exploratory data analysis showed significant variance in LOS for various procedures, so individual models could lead to better predictions.

Lastly, another avenue to explore is predicting total costs directly rather than estimating LOS to infer costs. My attempts to predict total charges so far haven’t yielded great results, with R² around 0.25. I believe that by implementing these improvements, we can build a machine learning model that delivers more accurate estimates for both LOS and associated costs, ultimately helping patients get the necessary insurance approvals.