Predicting Dialysis Center Star Ratings

Nichole Hartz

Capstone 2

Dialysis

- Needed in end stage kidney failure
- Performed in various settings
- Profit or Non-Profit
- Some chains

How does an individual's choice in dialysis center affect one's kidney health?

How do dialysis centers differentiate from one another?

The Problem

Data Information

Publisher: CMS

Time period: **01/01/2019 - 12/31/2019**

Facilities: All registered with Medicare

Number of dialysis centers: **7724**

Number of features: **119**

File format: csv

Dialysis Facility Compare Star Program

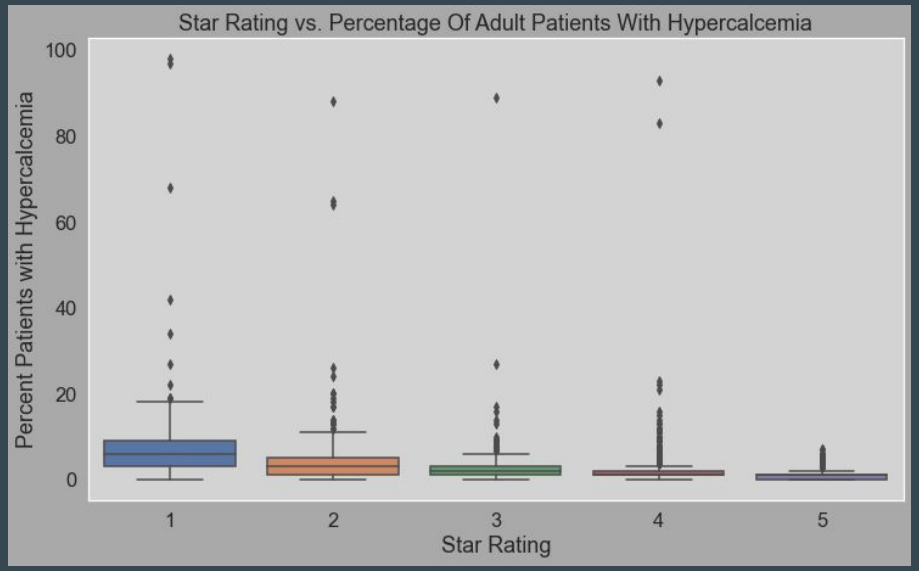
- Mortality ratios
- Hospitalizations
- Blood transfusions
- Incidents of hypercalcemia
- Percent waste removed
  - Hemodialysis
  - Peritoneal dialysis
- Percentage AV fistulas
- Percentage catheters (> 90 days)

Data Exploration

Health statistics & star rating

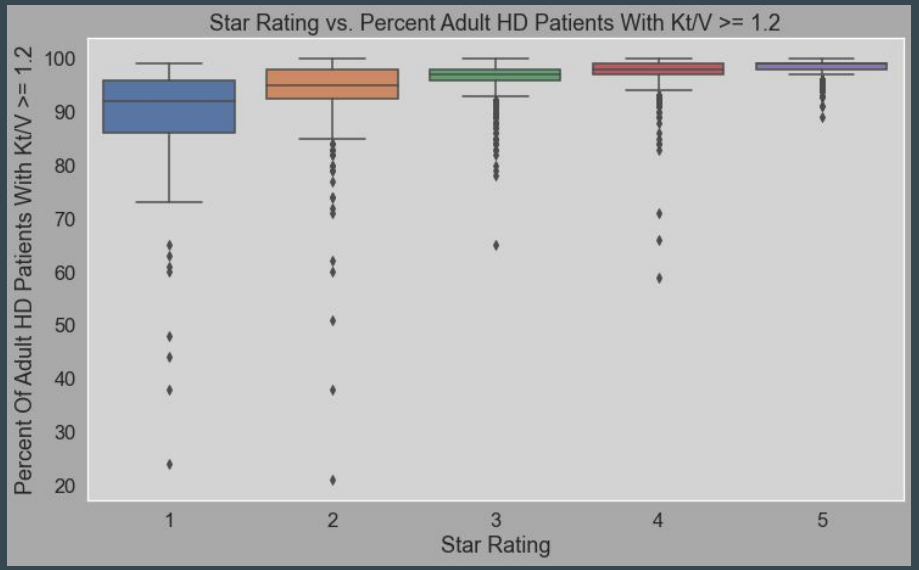
Between health statistics

Other feature to note



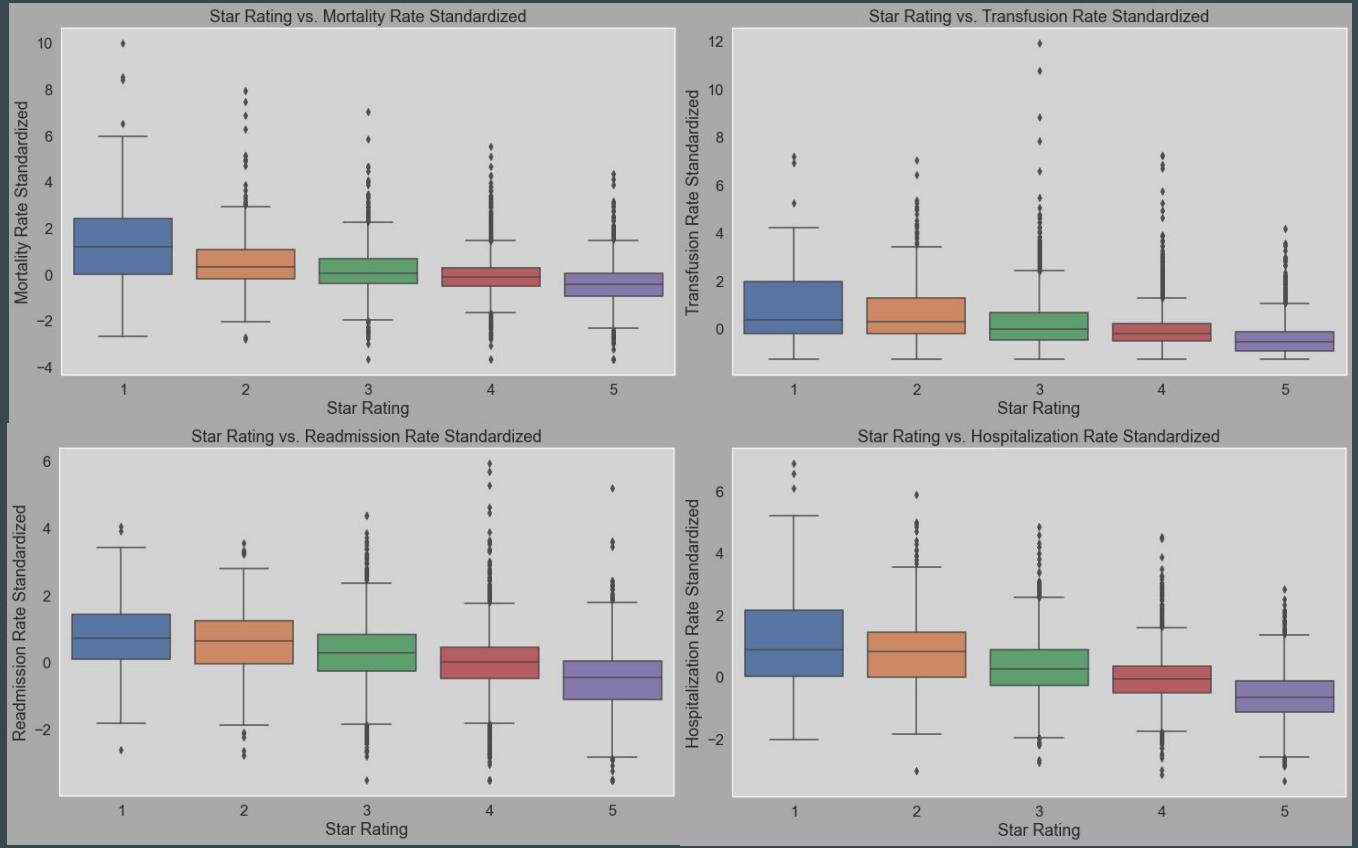
Smaller range for higher ratings

Blood measures of dialysis adequacy

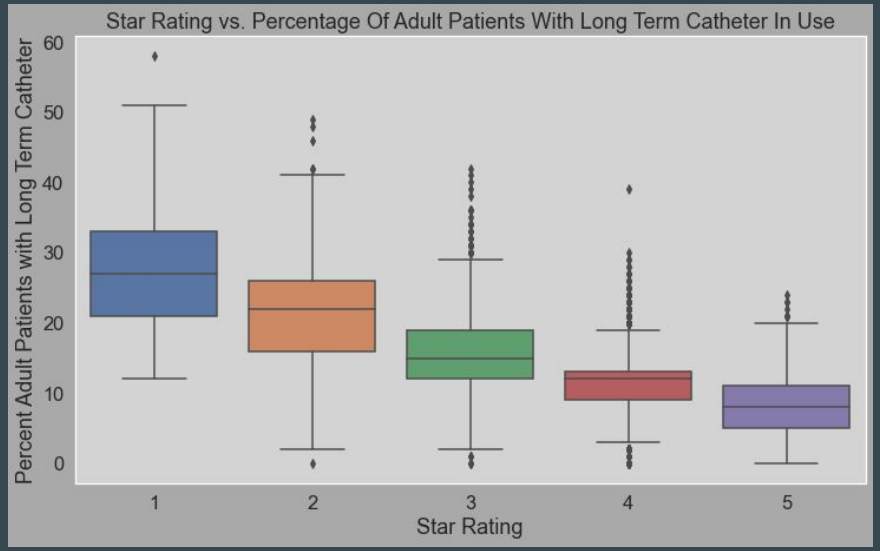


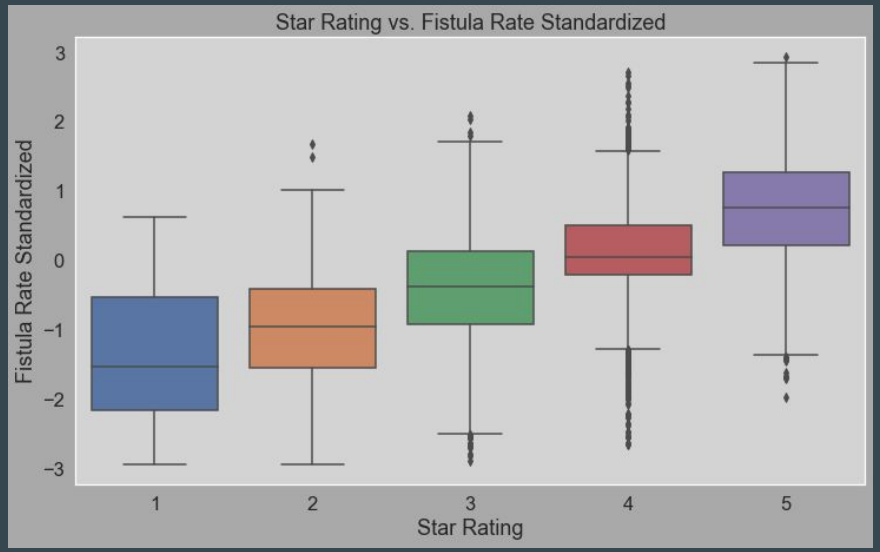
Narrower range = higher star rating

Multiple features show negative correlation with increased star rating



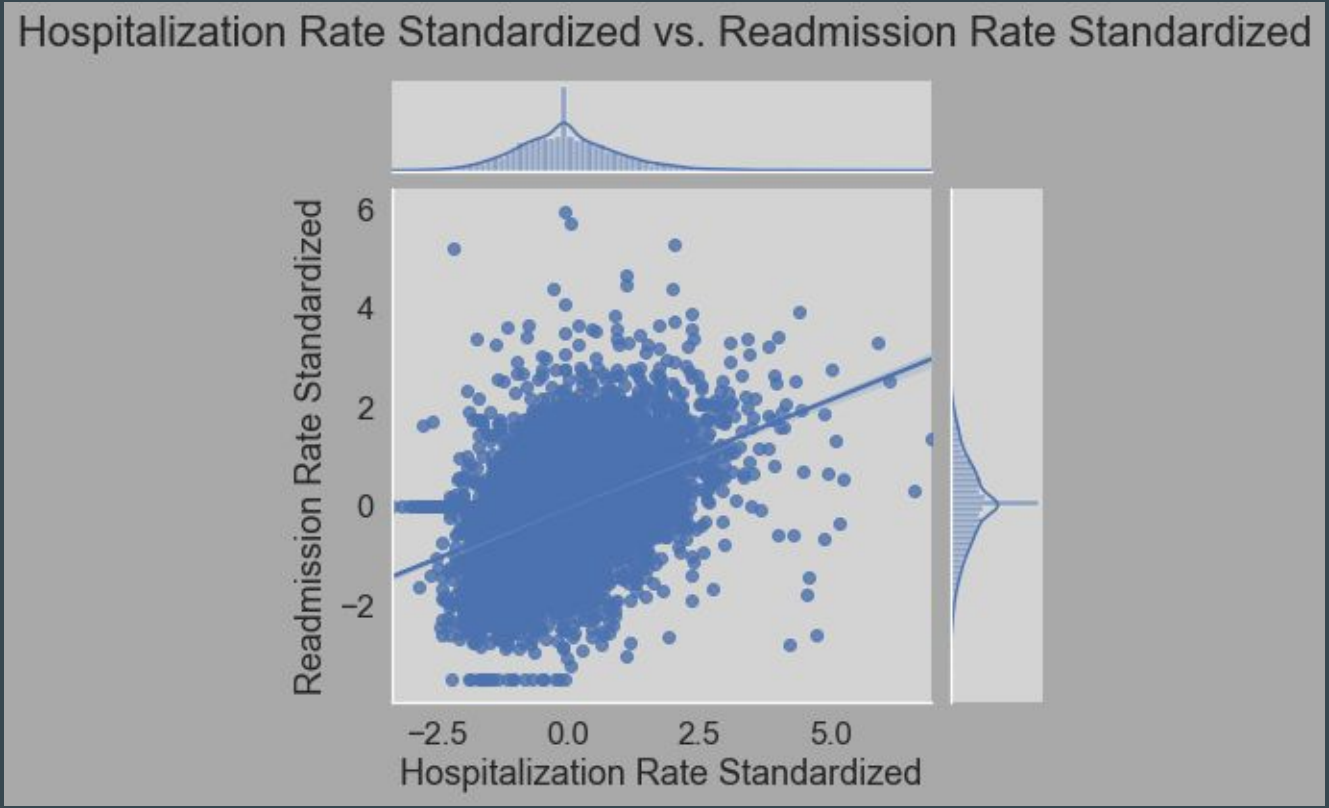
Fistulas & Catheters





Relatively strong correlations

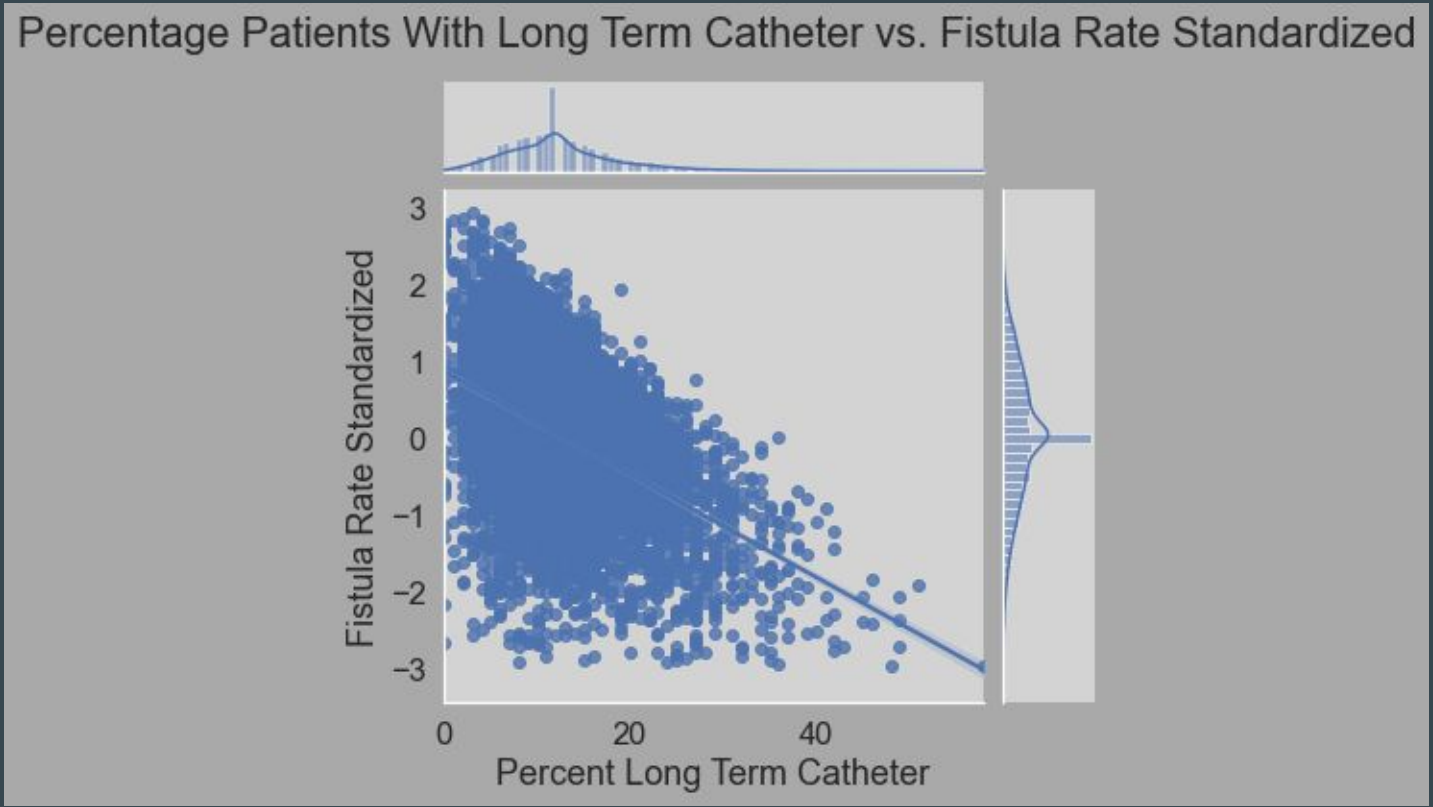
Hospitalization and readmission are positively correlated



May be a causing model to overfit

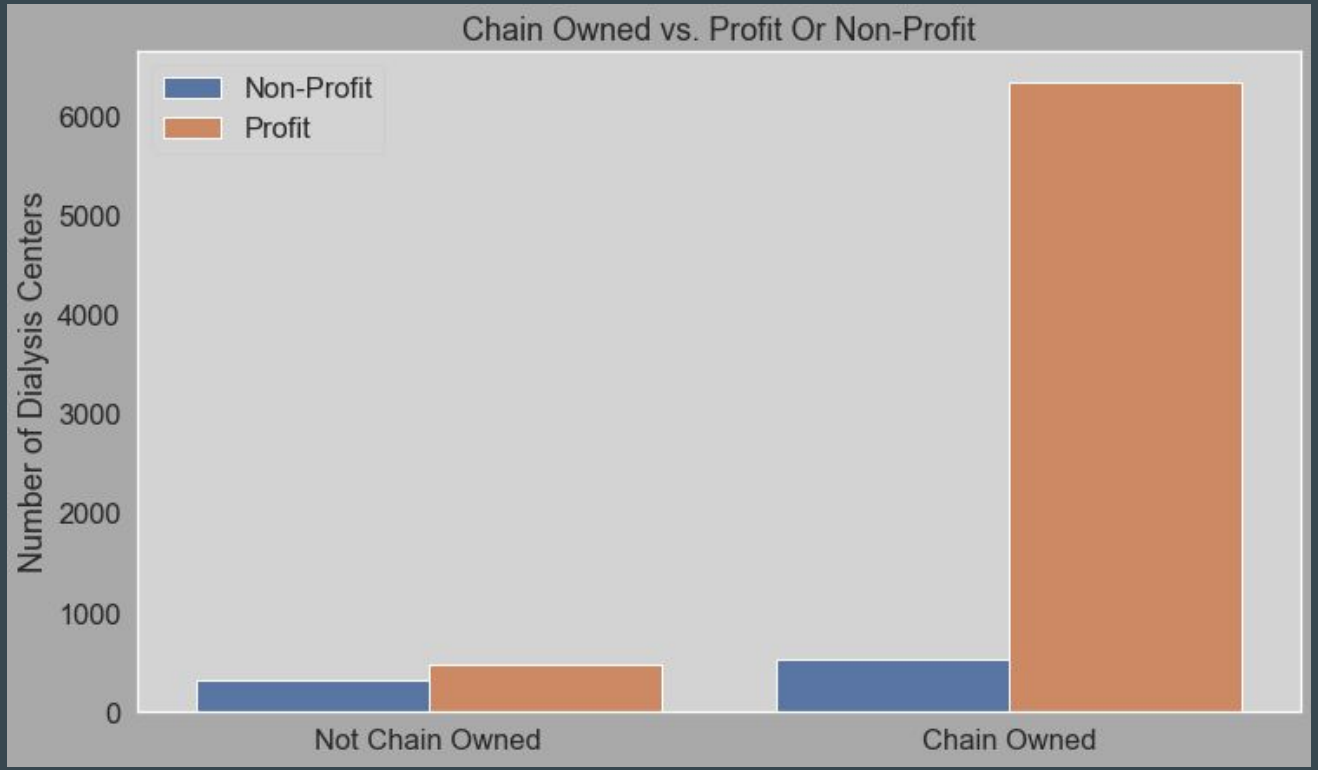
If in readmission already been in hospital

Catheter and fistula negatively correlated



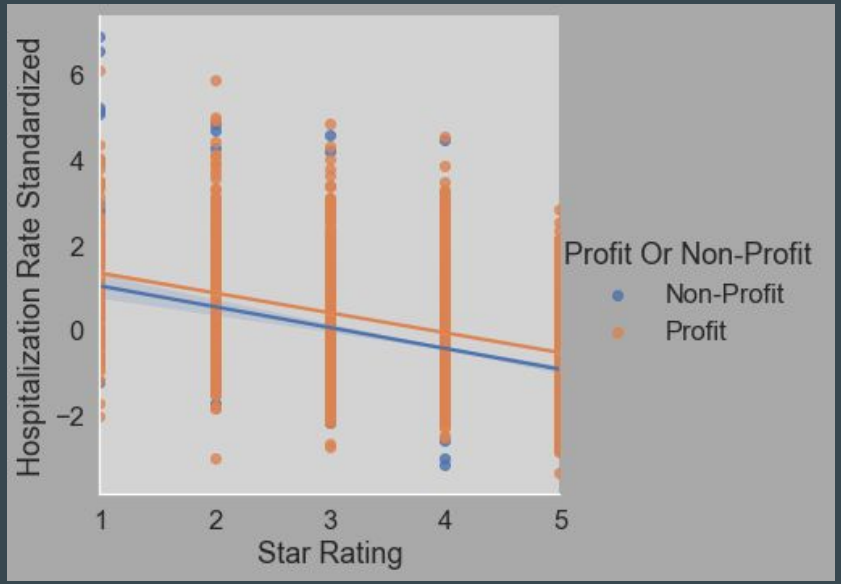
May be a causing model to overfit

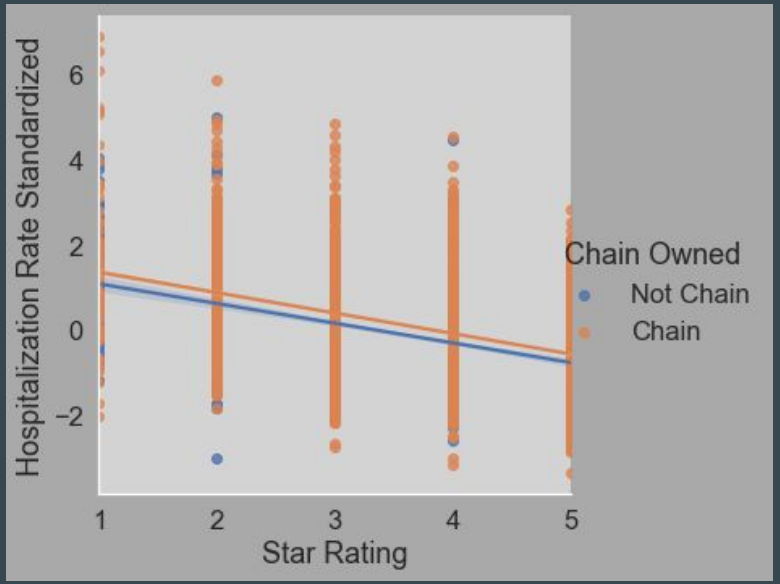
Patient typically only receives either a fistula or a catheter



Data Skewed? Majority of centers for-profit, chain-owned

Large overlap between profit and chain statuses





Model Overview

Supervised learning: Star Rating label

Multiclass classification: 5 star classes

Imbalanced: Majority of centers have 3, 4, or 5 stars

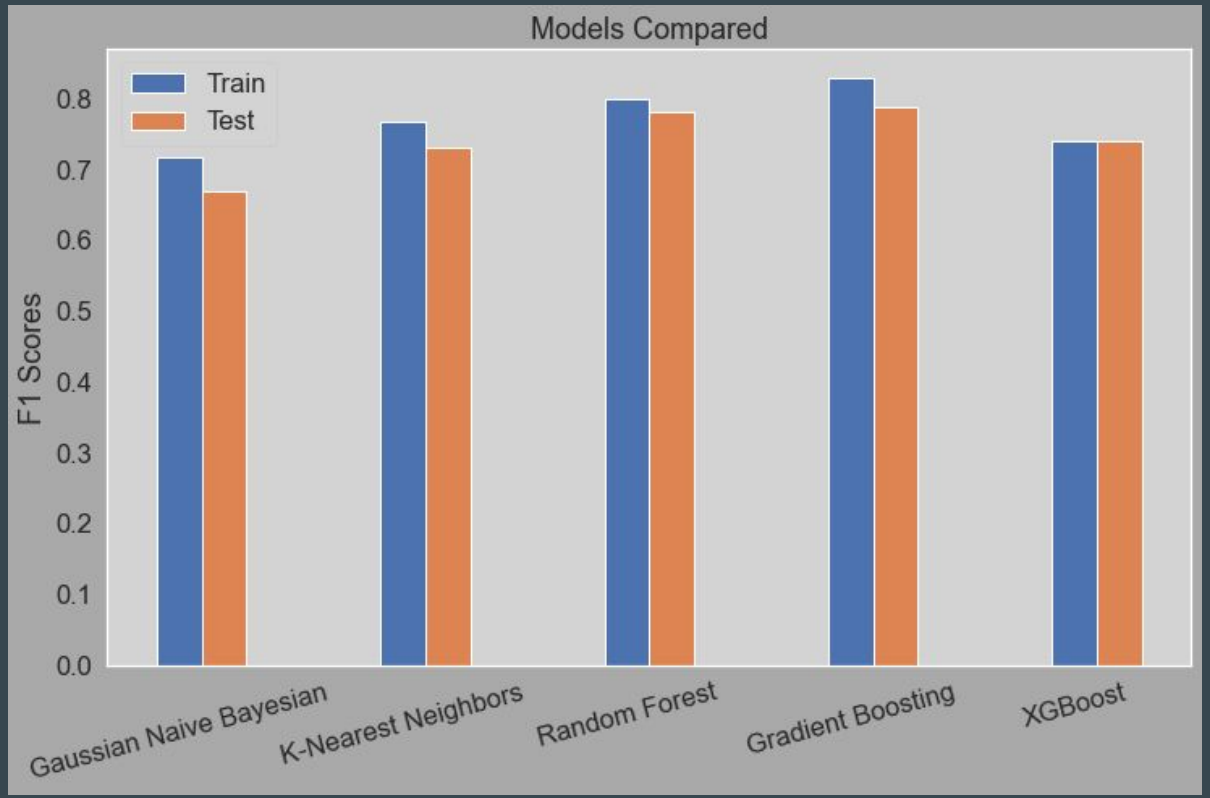
Tools: Scikit-learn, GXBoost

Modeling Steps

Preprocessing	Hyperparameter Tuning	Build Model	Cross Validate
• Label encoding • Standardize and remove outliers • Data split (75% train, 25% test)	• Parameters and ranges • F1 score with micro-average • Randomized Search with 5 cv	• Optimized parameters • Fit & predict model • Find runtime • Classification report	• Run cross validation on optimized model

Repeat for each model

Model Performance

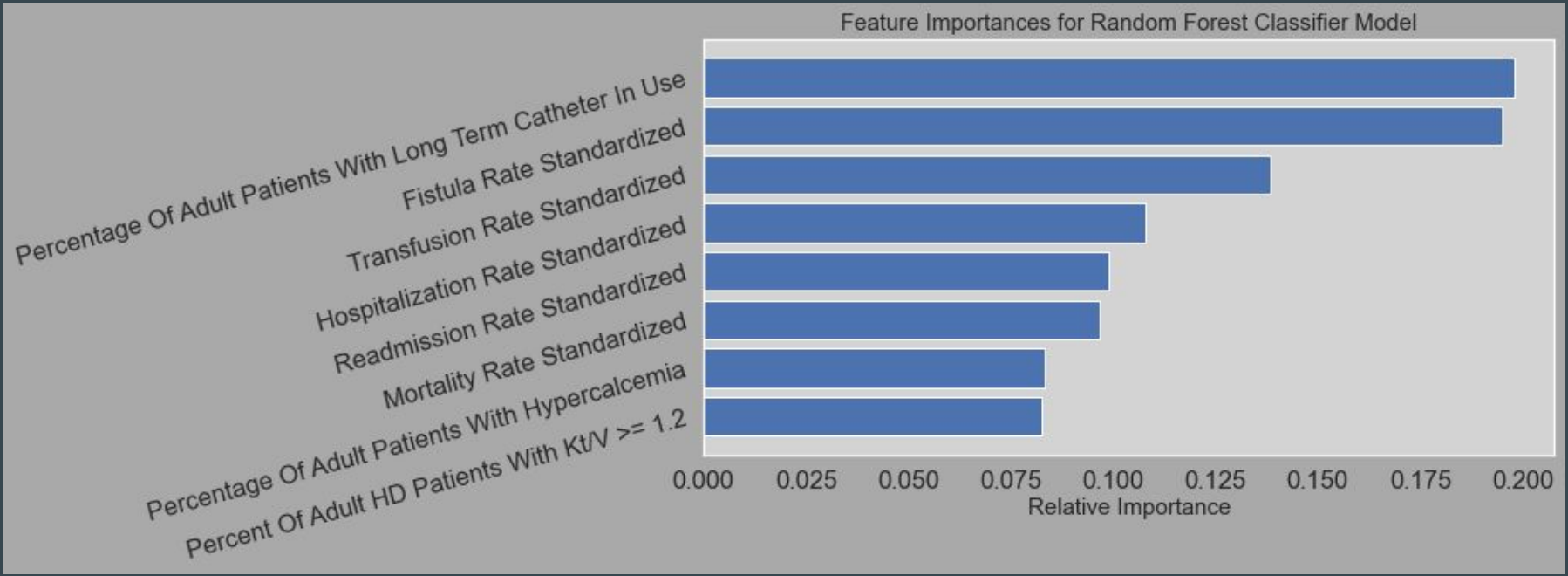


Random Forest and Gradient Boosting have similar scores

Random forest runs faster

Random Forest Best Model

Feature Importance

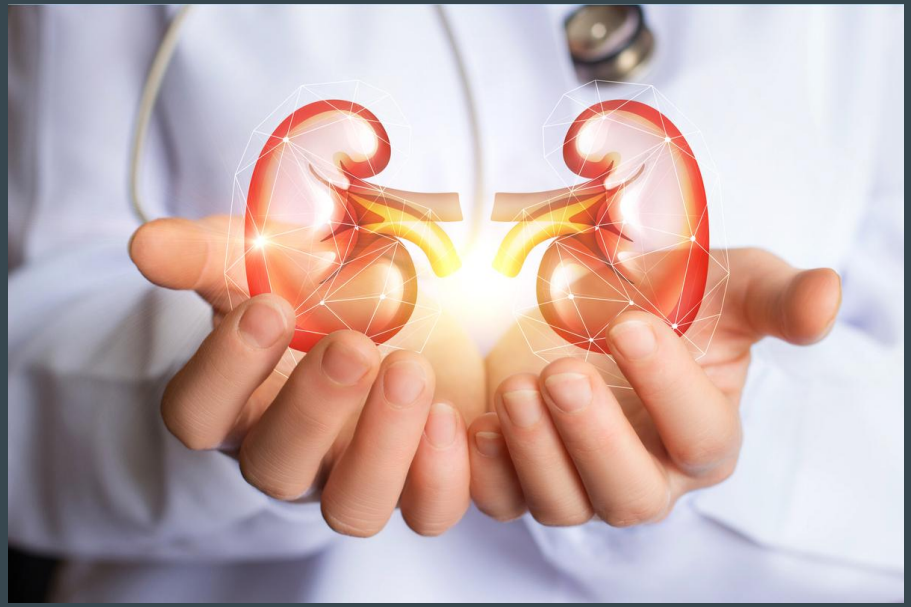


Constraints

- Only patients treated at dialysis centers
- Only centers registered by Medicare
- Health factors unrelated to dialysis
- Insufficient pediatric information
- Waste removed via dialysis broad term
- Overlap between health statistics

Ideas to Improve Model Performance

- Further hyperparameter tuning
- Find most important features in determining percent waste removed
- Gather pediatric patient information
- Directly link patient demographics to dialysis centers
- Find dialysis center explained reasoning for star rating



Conclusions

- Used 8 features (from star rating definition)
- Ran 5 supervised, multiclass classifiers
- Random Forest best model
- 75%/25% train/test split:
  - F1 test score = 0.802782
  - Runtime = 4.711022
- Modeling can be improved with further tuning & research