Hospital Stay Prediction

A Multi-Class Classification of Length of Hospitalizations
A Capstone Project By Hope Frost

How long will a patient be in the hospital? Will the hospital have enough resources?

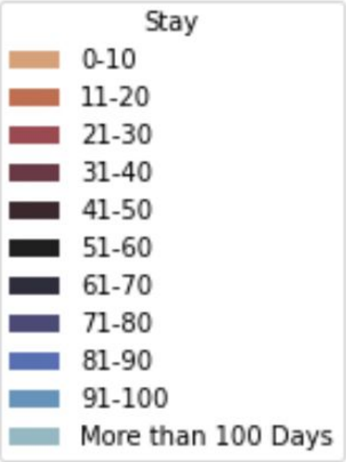
Hospitals are Short on Resources

If the hospital knows how long a patient will need care they can plan their use of resources to save stress on patients and staff; while allowing them to optimize the care and treatments given.

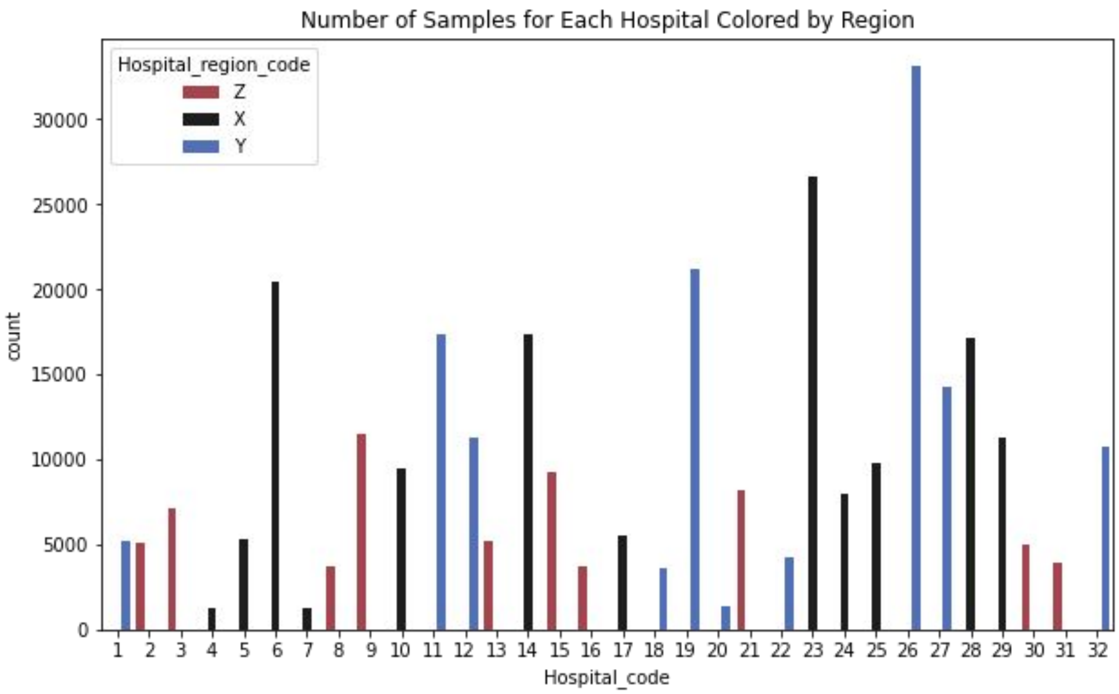
The Data

HealthMan a non-profit focused on the management and functioning of Hospitals, has posted a hackathon on Analytics Vidhya to assist healthcare management.

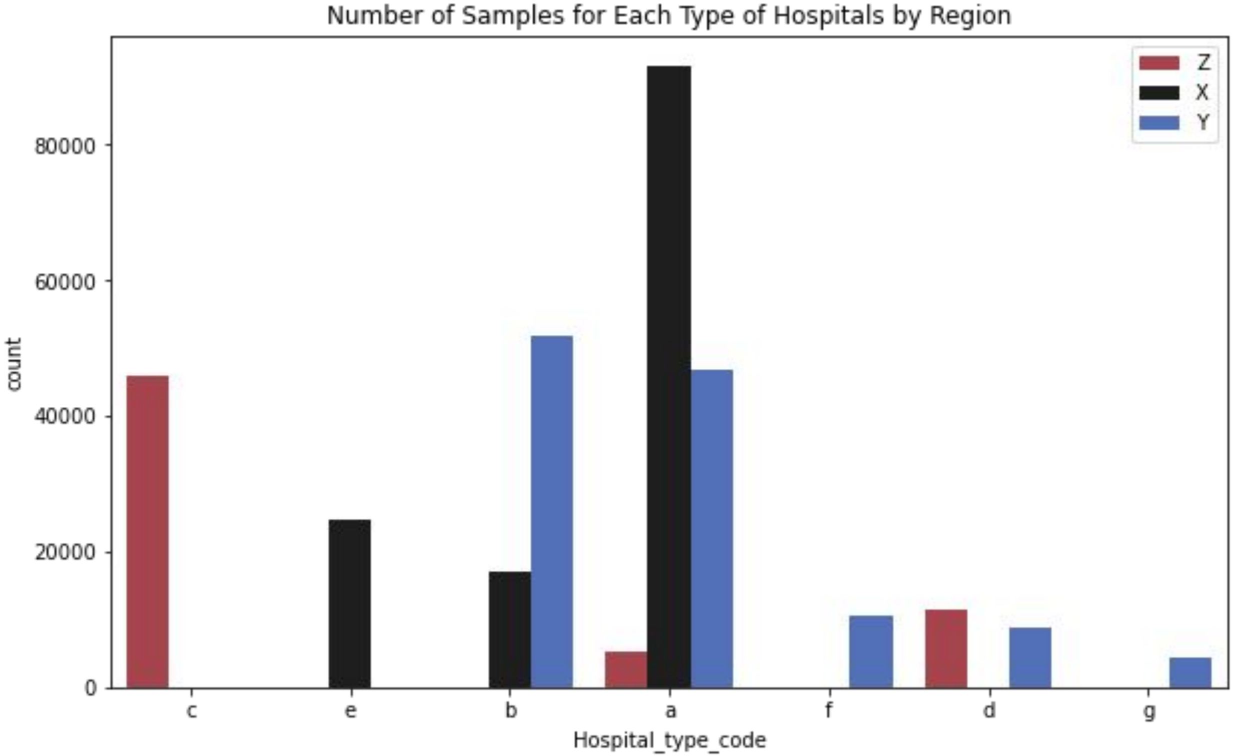
The train dataset consist of 318438 samples of admittance data with 18 features including the target feature of ten day ranges for the length of time spent in the hospital.

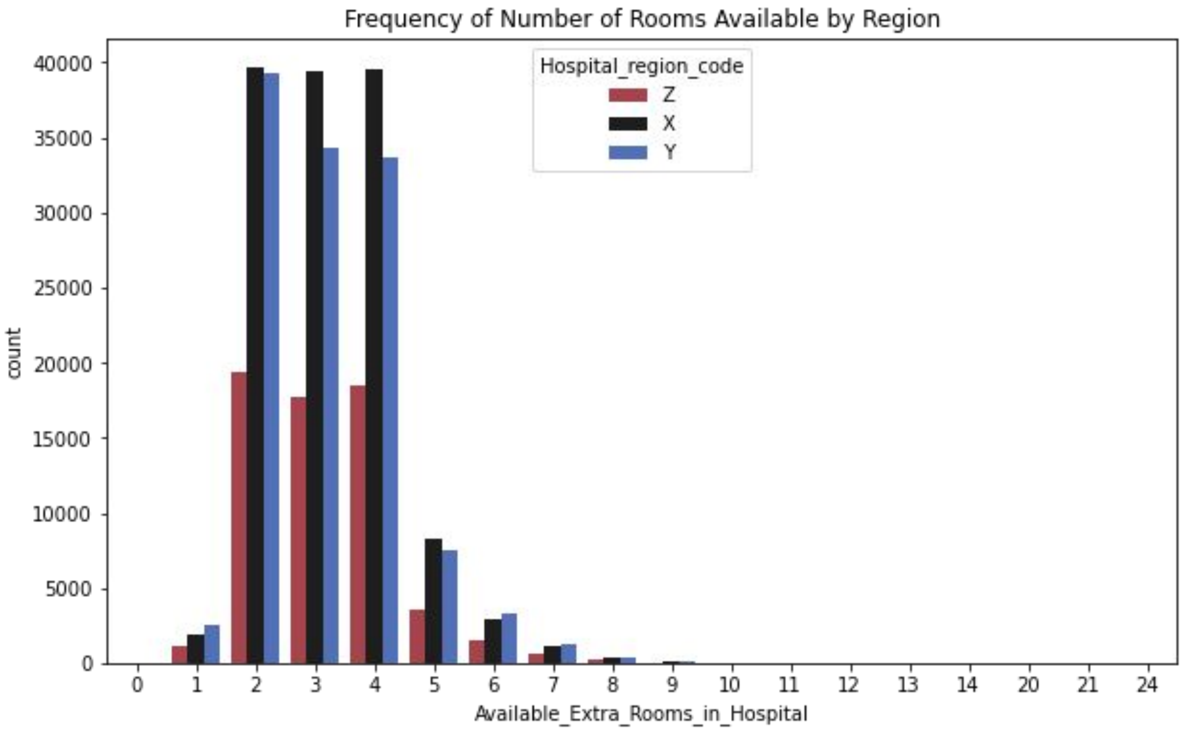


- 32 hospitals
- 3 different regions
- A number of descriptive features for the hospitals
- A couple for the patients admission assessment



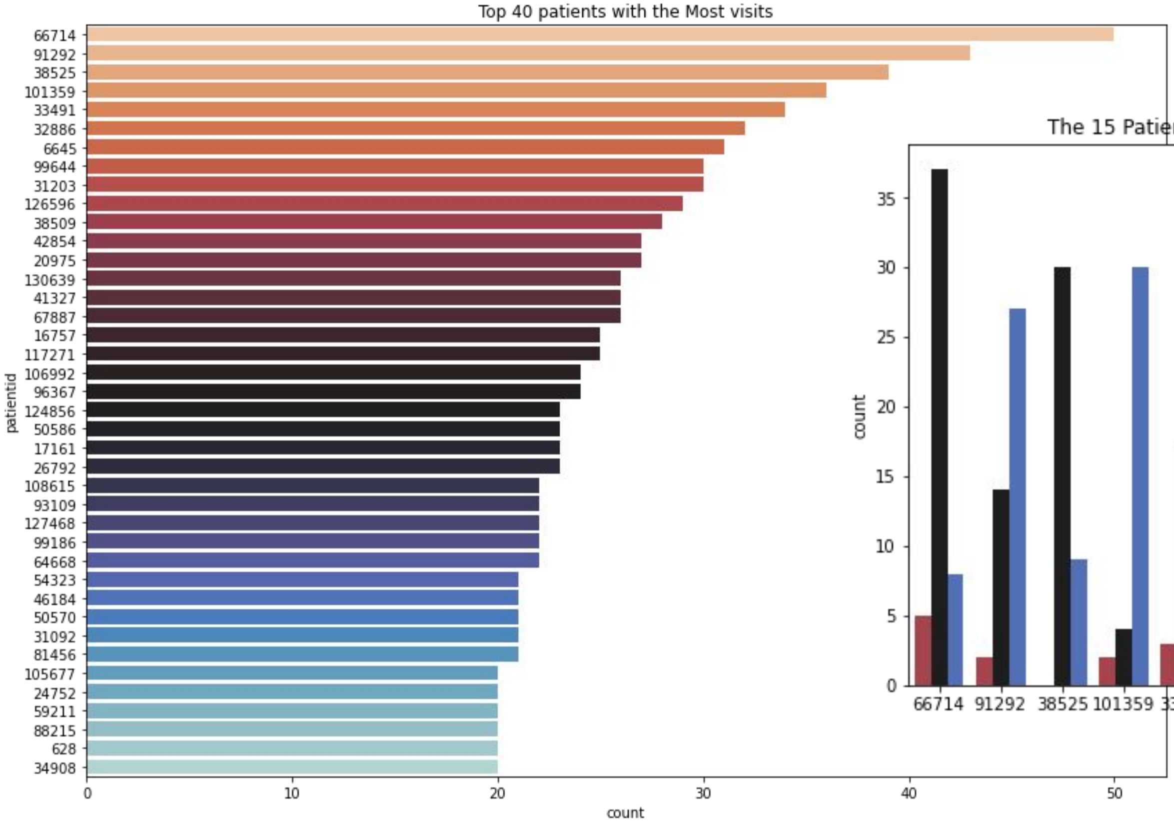
For privacy the dataset has been manipulated making it hard to interpret.

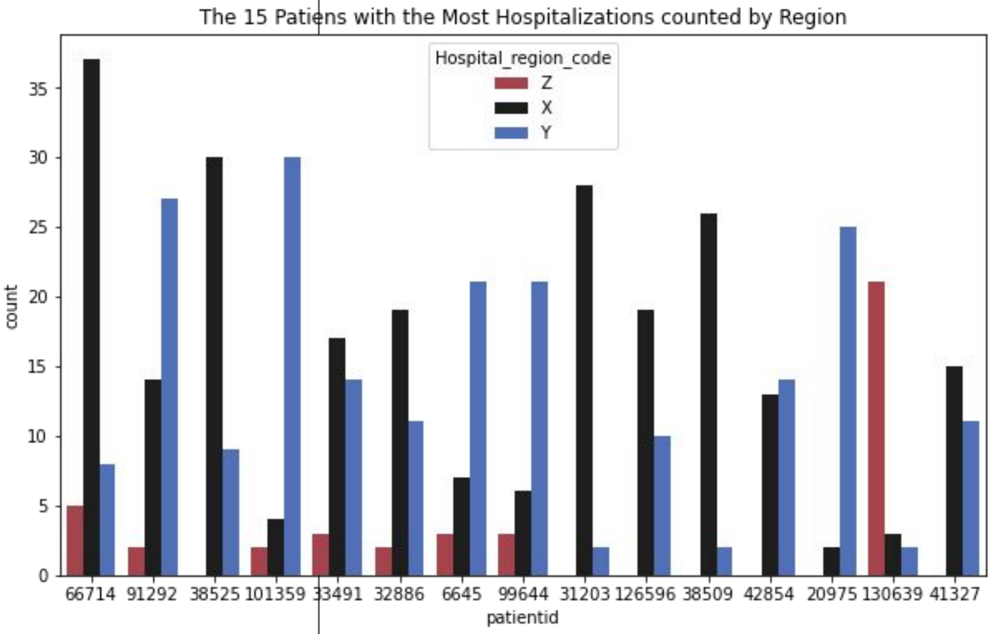




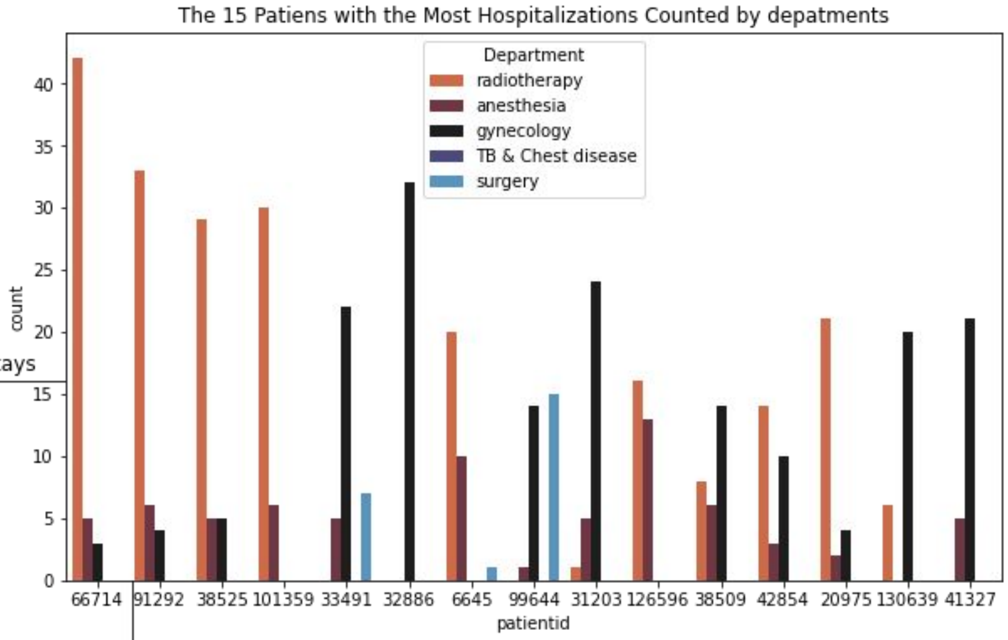
The volume of patients in each region indicates that region Y has larger hospitals with more capacity.

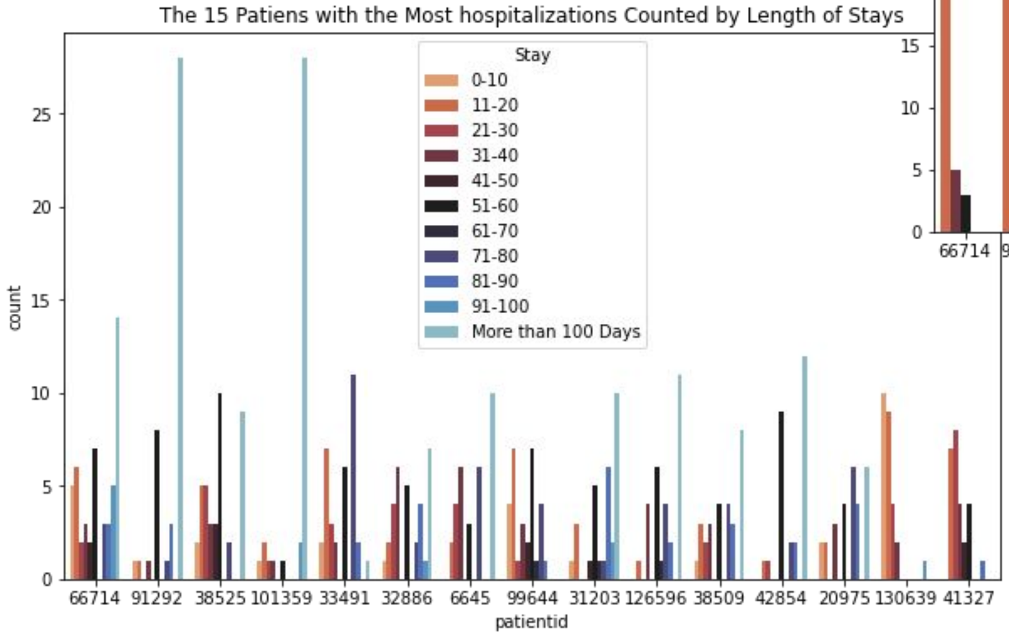
Trends for Patients with Multiple Hospitalizations





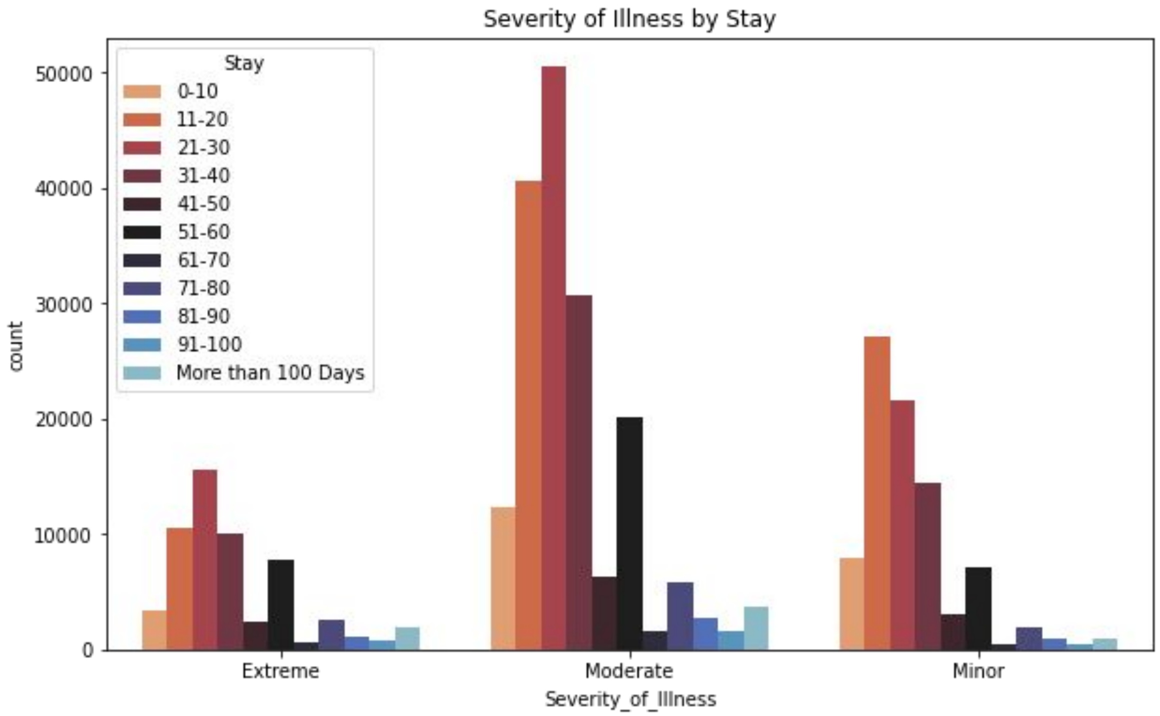
At least 8 years of records possibly more if spread out





Some very sick patients rather than data issues

Classification Ranges Not Equally Distributed



Features Engineering

Count Features

Impute a count of how many samples have that feature from the train set.
- 9 count features

Feature by Feature

Impute counts of occurrences of a feature by another feature. Count of rooms available, bed grade or visitors by location, type or department of the hospital.
-21 feature by feature counts

Patient Traveled

If the patient traveled to get to the hospital it may increase the time needed to recover enough to get home. Create a binary feature with one representing that the patient lives in a different region then the hospital.
- 1 feature

Preprocessing and Modeling

- Split the original train set into train and test sets
- Engineered features off train set only
- One hot encode the categorical features
- Encoded target feature for Multi-Classification.
- Created Random Forest model
- Tuned Hyperparameters

Metrics

The Analytics Vidhya hackathon score is based off 100% accuracy.

At of the time the contest closed the highest accuracy achieved was a 43.908% with 19638 submissions entered during the time of the contest.

Model

The Random Forest model predicts 0 test samples in the two classification ranges we expected it to struggle with.

69.65% of predictions are within one range off

Accuracy

Train score:
52.92009312140955%

Test score:
42.10275091068961%

Hackathon public score:
42.0013863047682%

Hackathon private score:
41.7048701890922%

Top 1.5%

Score Rank 266

Real World Value:

While a 42% accuracy may seem low, almost 70% of predictions are within a 20 day range. The model is producing a ballpark for hospitals to plan off.

Further Steps

To increase our accuracy additional information is needed. Date of admission, or at the least age at time of admission would be very helpful. Information on regions such as population and size of hospitals would also be beneficial

Standardizing and scaling with a pipeline to try other models is an option. However, without better data a model is unlikely to perform much better.

Thank You

A special thanks to Silvia Seceleanu
And to Springboard