Healthcare Management Analytics

By: Anthony Huang, Vivek Mehendiratta, Rhiannon Pytlak, Vishal Gupta

Analysis Outline

Presentation Overview


Problem Statement


Data Description


Project Objective


Exploratory Analysis Insights


Hypothesis


Feature Engineering


Model Selection Methodology


Cross Validation


Hyperparameter Tuning


Feature Importance

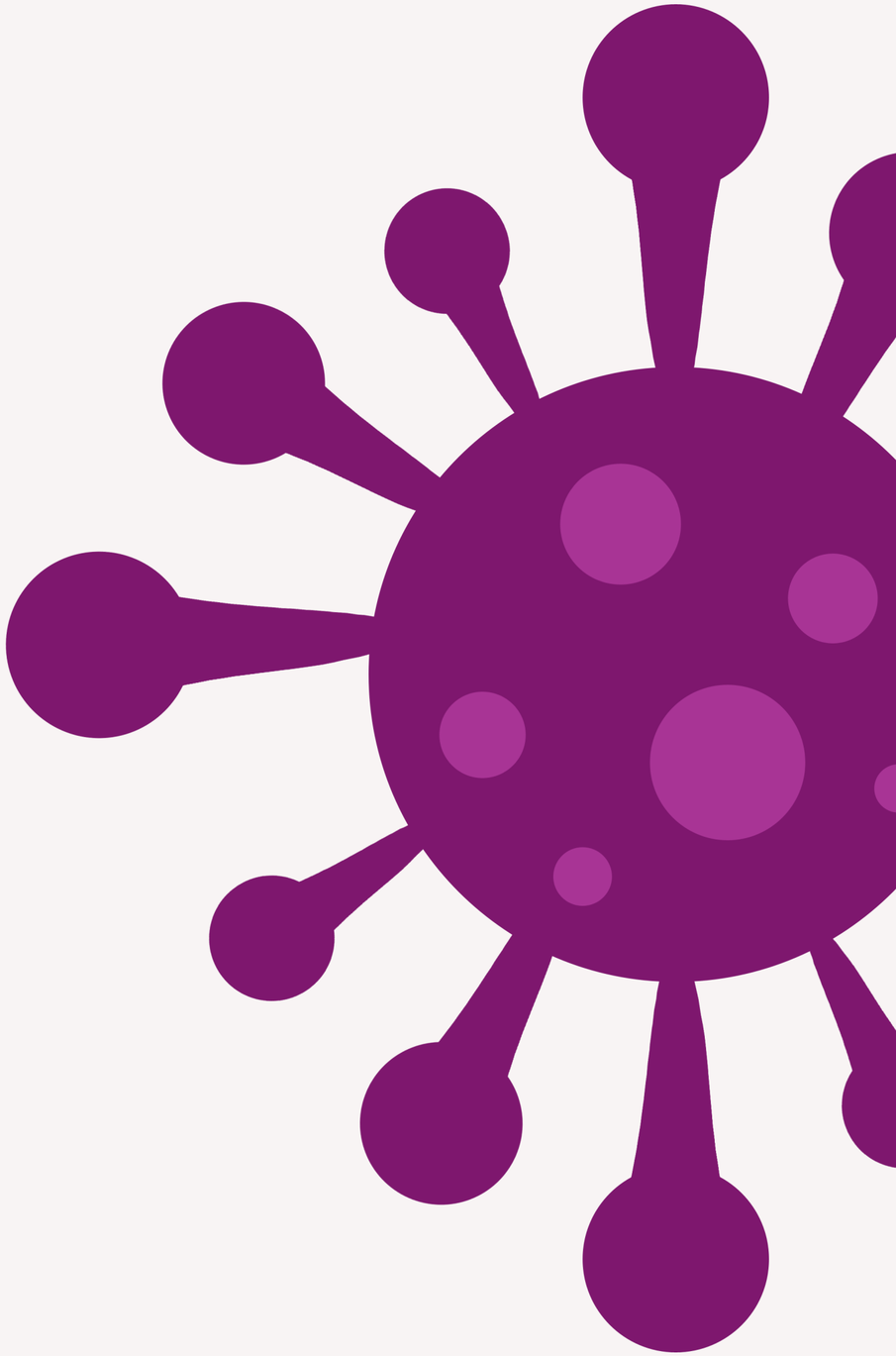

Conclusion

Problem Statement

The COVID-19 pandemic has proven that efficient healthcare management is more critical now than ever.

Since the pandemic began, more and more people have been flocking to healthcare facilities, making it extremely tough for hospitals to manage patient case loads.

If healthcare officials were able to accurately predict how long a certain patient would be staying in the hospital, the hospital could serve more patients.



Healthcare Data Description

- Data source: Kaggle
- Contains: patient attributes & length of stay at hospital
- Target variable: 'Stay' is divided into 11 classes ranging from 0-10 days to 100+ days

	Column	Description
0	case_id	Case_ID registered in Hospital
1	Hospital_code	Unique code for the Hospital
2	Hospital_type_code	Unique code for the type of Hospital
3	City_Code_Hospital	City Code of the Hospital
4	Hospital_region_code	Region Code of the Hospital
5	Available Extra Rooms in Hospital	Number of Extra rooms available in the Hospital
6	Department	Department overlooking the case
7	Ward_Type	Code for the Ward type
8	Ward_Facility_Code	Code for the Ward Facility
9	Bed Grade	Condition of Bed in the Ward
10	patientid	Unique Patient Id
11	City_Code_Patient	City Code for the patient
12	Type of Admission	Admission Type registered by the Hospital
13	Severity of Illness	Severity of the illness recorded at the time o...
14	Visitors with Patient	Number of Visitors with the patient
15	Age	Age of the patient
16	Admission_Deposit	Deposit at the Admission Time
17	Stay	Stay Days by the patient

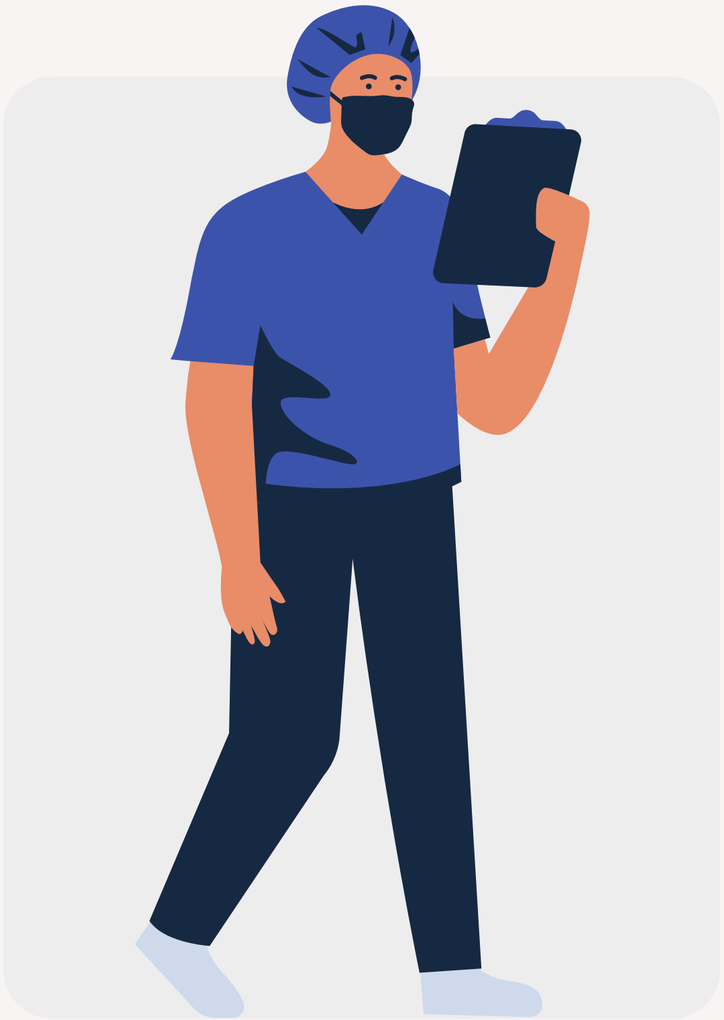
Objective & Procedure

What do we want to do?

Our objective is to predict the length of stay for each patient on a case-by-case basis so that hospitals can use this information for optimal resource allocation.

How?

Using past patient data, we will attempt to construct a machine learning algorithm that can predict how long a patient would require hospitalization.

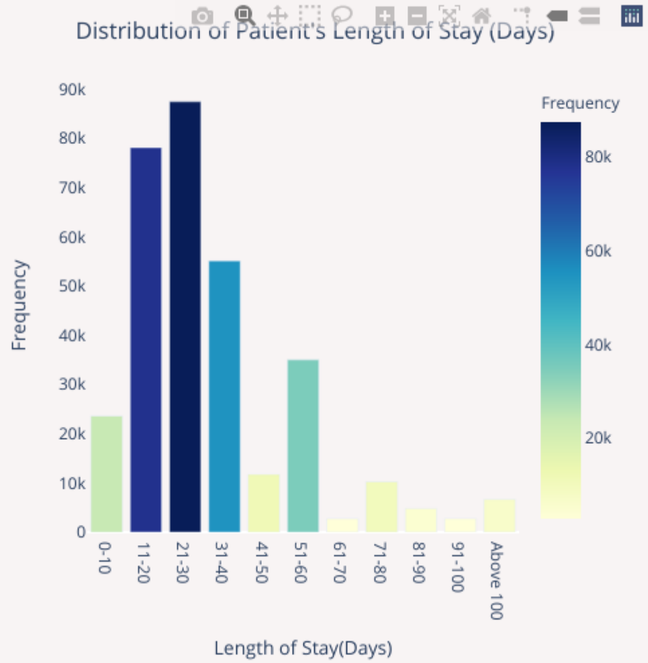


Exploratory Data Analysis Insights

Distribution of Target Variable

- A majority of admitted patients (76%) tend to stay hospitalized anywhere between 0-40 days.

- When we plotted the Stay variable in a histogram, we found a **multimodal distribution**



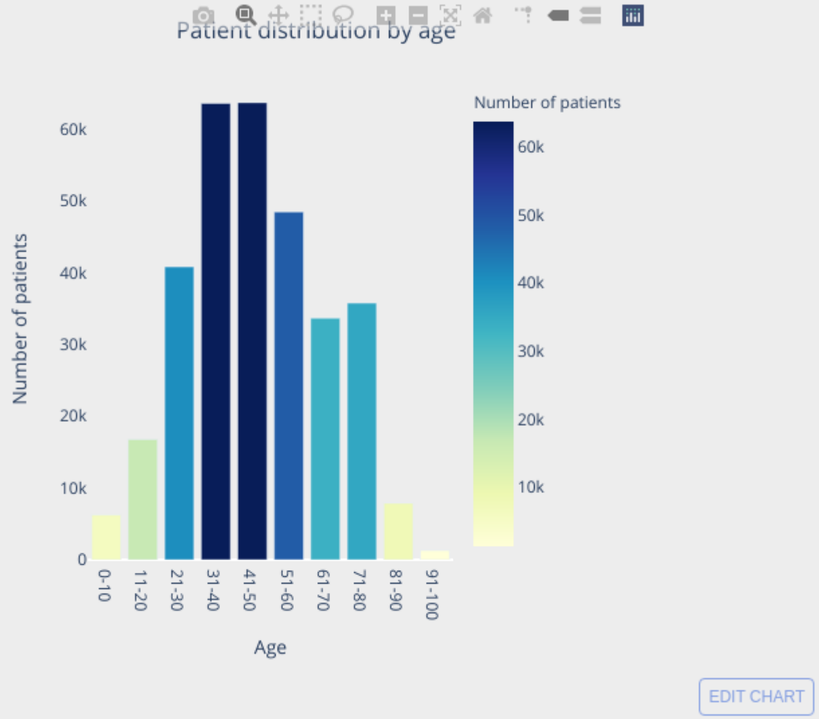
EDIT CHART

Positive Trend: Stay & Visitors

- On average, a patient will be admitted to the hospital with 2 visitors.

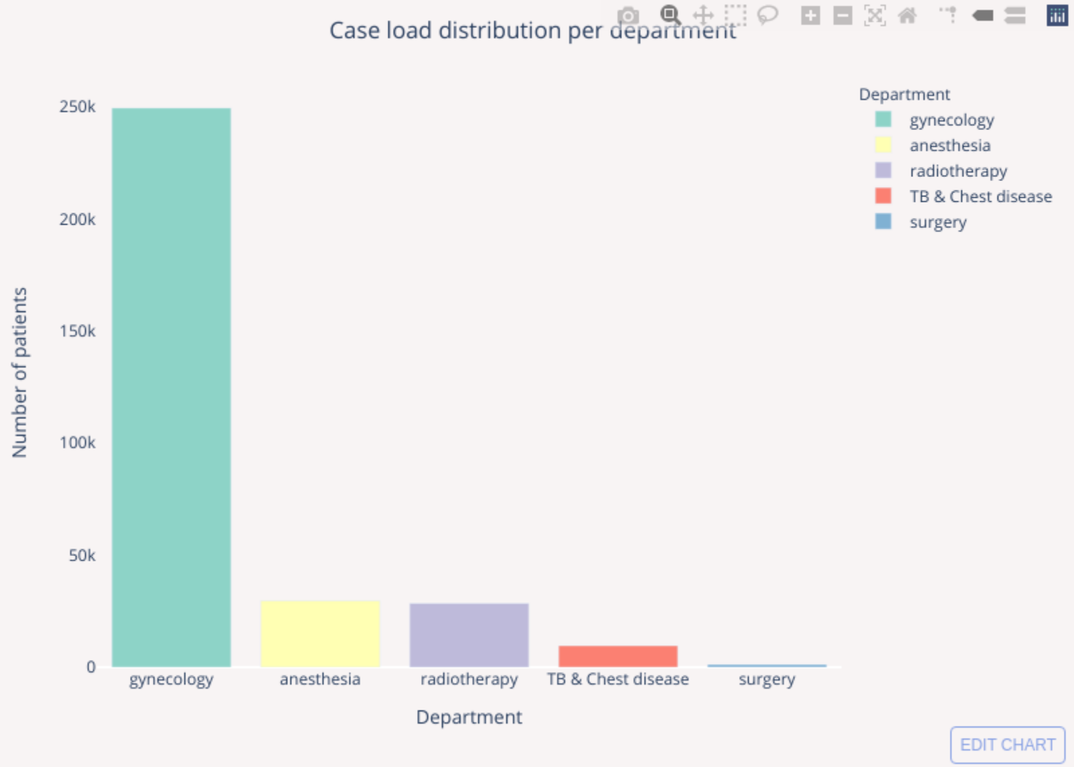
- There seems to be a positive trend between length of stay and the number of visitors that a patient is admitted with

Number of Hospitalized Patients by Age



A majority of hospitalized patients fall into the **31-40** and **41-50** age buckets.

Number of Hospitalized Patients by Department



80% of patient admission is for the Gynecology department, while the surgery department has the smallest number of caseloads

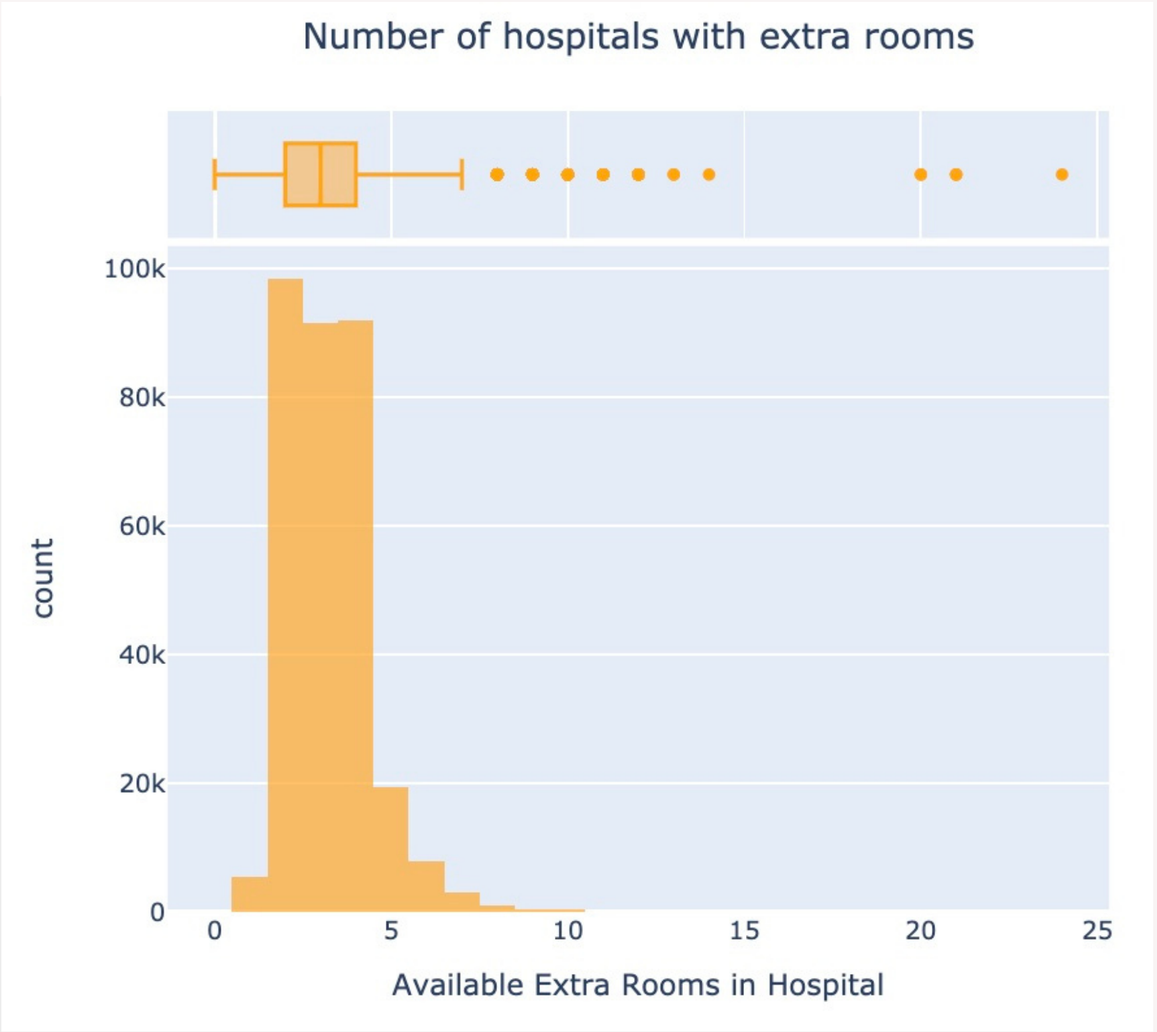
Stay	max_deposit	min_deposit	average_deposit	median_deposit
0-10	9673.0	1801.0	4615.214625	4513.0
11-20	10419.0	1832.0	4931.124829	4800.0
21-30	10729.0	1807.0	5025.310329	4886.0
31-40	11008.0	1820.0	4871.071067	4708.0
41-50	11008.0	1825.0	4888.818530	4803.0
51-60	10771.0	1831.0	4748.784397	4526.0
61-70	10254.0	1809.0	4845.449344	4768.0
71-80	10842.0	1833.0	4709.845426	4500.0
81-90	10729.0	1827.0	4590.644688	4291.0
91-100	10506.0	1805.0	4715.538879	4530.0
Above 100	10999.0	1800.0	4649.341763	4384.0

Admission Deposit Variable

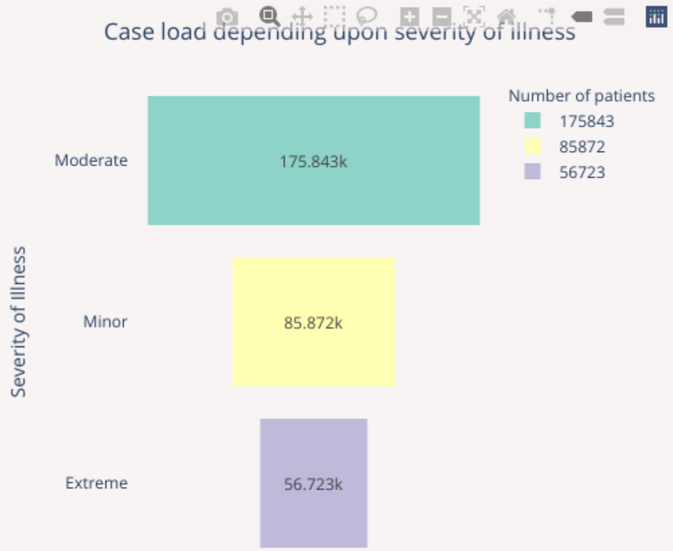
- On average, a deposit of 4,880 is paid before admission of the patient.
- There is no information available about where this data was collected geographically

Rooms Available at Time of Patient Admittance

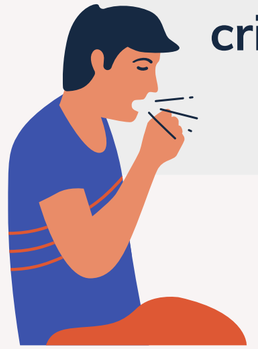
When a patient is being admitted to a hospital, there is a high probability (>90%) that there are 2-4 extra rooms available in that hospital at that moment.



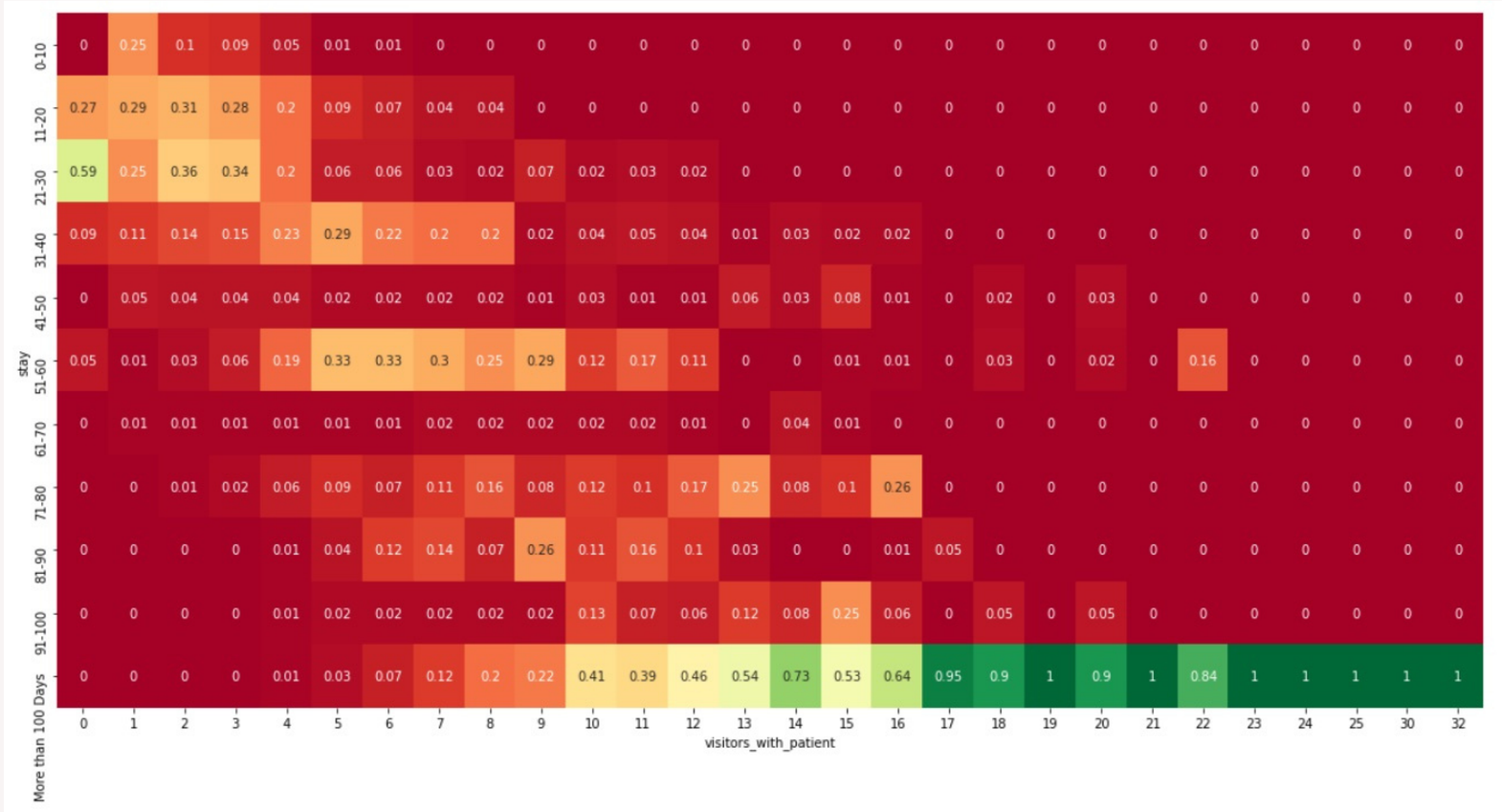
Number of patients in each illness severity level



- More than half of the patients are admitted with moderate illness severity
- 18% are in extremely critical status



EDIT CHART



Visitors with Patient (at time of admission) Variable Insight

When there is a higher number of visitors present during admission, the conditional probability of a patient being hospitalized for a longer time period increases.

Hypothesis

Based upon what we've seen so far, we hypothesized that the following variables would have an effect on a patient's length of stay

Age

Younger patients may recover faster

Admission Deposit

Possible indicator of patient's financial stability

Visitors with Patient (at time of admission)

Could indicate illness severity, therefore increasing hospitalization time

Department

Maybe a patient's length of stay will vary based upon what department they're in?

Severity of Illness

More severe illnesses might lead to a longer hospitalization period

Feature Engineering

Removed Variables

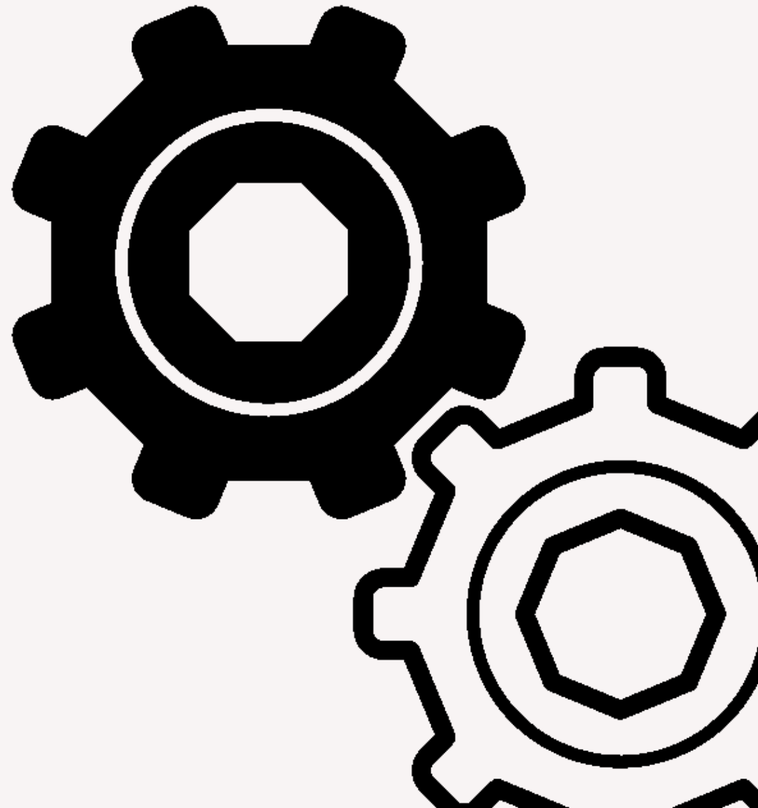
- Case ID
- Patient ID
- City Code Patient

Categorical Feature Encoding

- Hospital Type Code
- Ward Facility Code
- Ward Type
- Type of Admission
- Severity of Illness
- Type of Admission

Numerical Data Scaling

- Available Extra Rooms in Hospital
- Visitors with Patient
- Admission Deposit



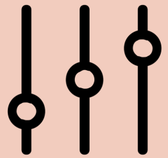
Model Selection Methodology

How are we measuring accuracy?

What models are being used?

How are we doing cross-validation?

Which parameters are hyper-tuned?



Out-of-Sample Prediction Accuracy by Model

Method: K-fold Cross-Validation (5-fold)

Decision Tree
28.7%

Naive Bayes
32.5%

Random Forest
33.4%

Logistic Regression
39.8%

Hyperparameter Tuning

Logistic Regression

Search Method: Grid Search

Parameters: max_iter, C, penalty

Accuracy Score: 39.8%

Random Forest

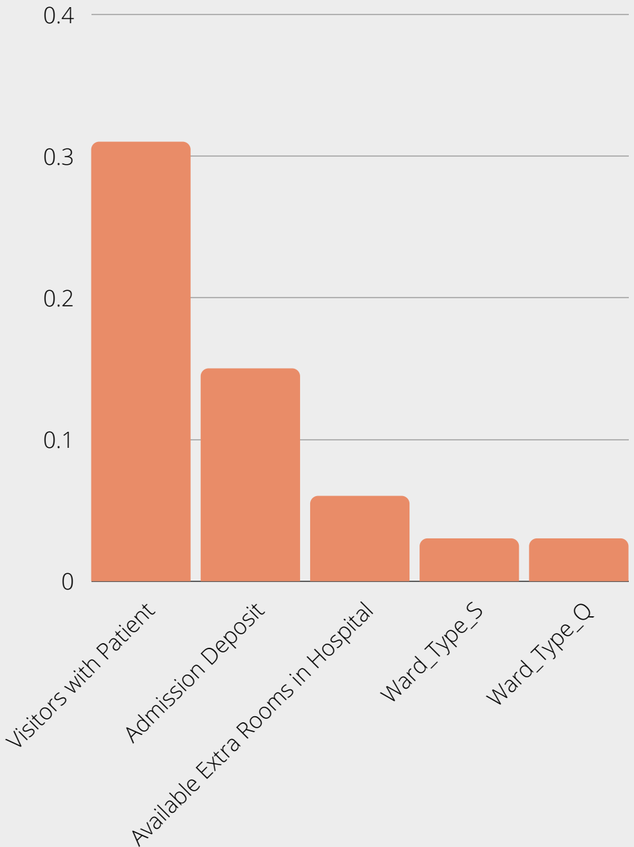
Search Method: Random Search

Parameters: max_depth, max_features, min_samples_leaf, min_samples_split, n_estimators

Accuracy Score: 40.8%

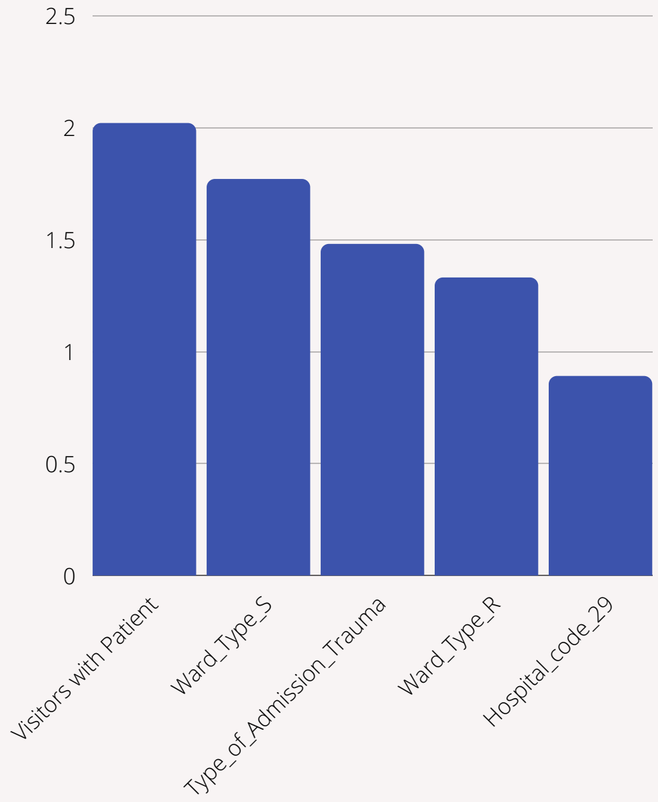
Feature Importance

Logistic Regression



1. Visitor with Patient (-)
2. Admission Deposit
3. Available Extra Rooms

Random Forest



1. Visitor with Patient
2. Ward Type
3. Type of Admission

Findings Summary

Best Prediction Model

The best out-of-sample prediction accuracy achieved was 40.8% with the hyperparameter-tuned **random forest model.**

Limitations

- We speculate that there are external factors not being taken into account.
- There are innumerable variables that could impact the duration of any given patient's hospital stay.
- There is not a **feasible** way to capture all of those variables for every patient and produce an interpretable prediction model for their hospitalization period.

Questions?