Hospital Length of Stay Prediction

By Ketaki Nagarkar
Springboard Data Science Capstone
Mentor: Ben Bell

Background:

While working in the hospital in the acute care/post-operative care floor, date of discharge was a question that often came up.

A patient's family member would inquire: *'we traveled from out of town, how long will my husband be admitted, I have to plan my work leave, hotel stay and transportation accordingly and of-course budget in the medical expenses'.*

The floor manager would check to see *'We have two patients awaiting elective surgery, and one more awaiting emergency admission. Do you know when a hospital bed will be available on this floor?'* and *'I also need to know how many patients will need overnight stays so I can allocate adequate nursing staff to ensure good quality of care'*.

How many days Length of Stay(LOS)? Or When is the patient going to be discharged? is not a question that is not easily answered as it is dependent on a number of factors. The goal of my project is to be able to predict LOS following patient admission and also explore which admission factors influence variability in LOS.

Why is predicting Length of Stay important ?

Predicting patient Length of Stay (LOS) at admission will assist with

- Improved resource allocation for medical facilities
- Plan medical expenditure for patients and hospitals
- Improving quality of medical care for patients

Who will benefit from this project?

- Hospital management
- Insurance Companies/ Primary Insurance Payers
- Patients and their families

Dataset:

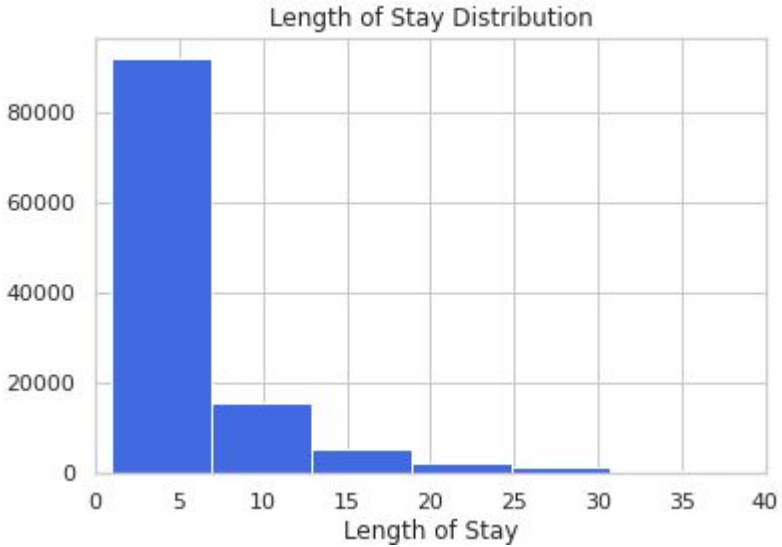
De-identified healthcare data was obtained from New York State Department of Health. A 5 percent random sample of this dataset was used for the project.

Hospital Inpatient Discharges (SPARCS De-Identified) Downloadable File: 2014 | State of New York (ny.gov)

The data consisted of hospital stay details including diagnosis, procedures, diagnosis and procedure codes, patient characteristics, length of stay and charges

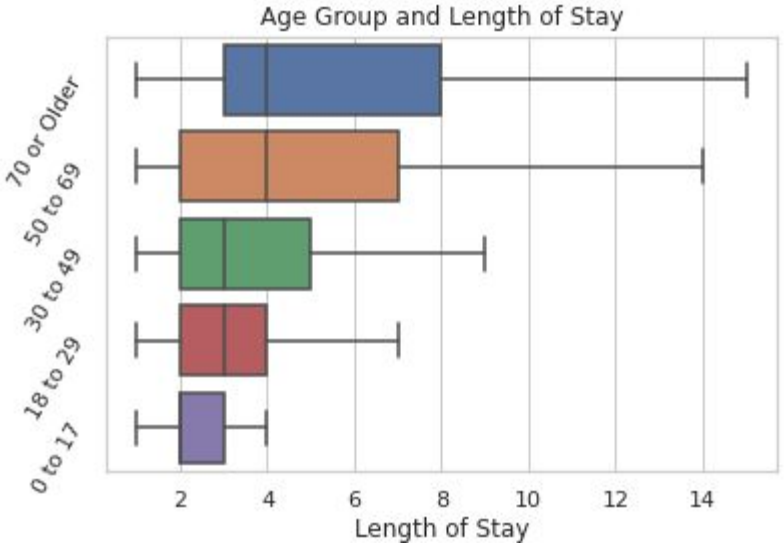
The data was checked for missing values and inconsistencies prior to EDA. Abortion records were redacted in this datset and corresponding values for columns were null, these records were dropped.

EDA: Analysis of Length of Stay (LOS)



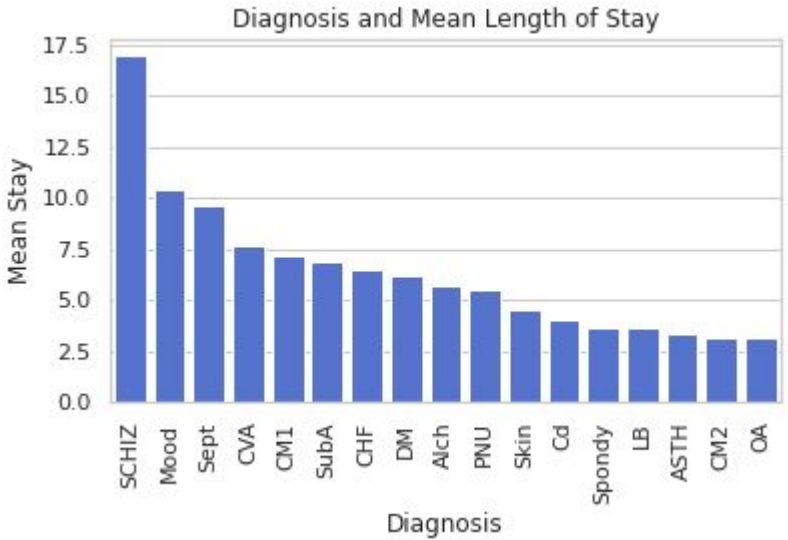
A zoom in shows that majority of hospital stays lie at or below mean i.e. 6 days or less. A smaller fraction of the hospital stays are more than mean.

EDA: Age appears to be positively correlated with LOS



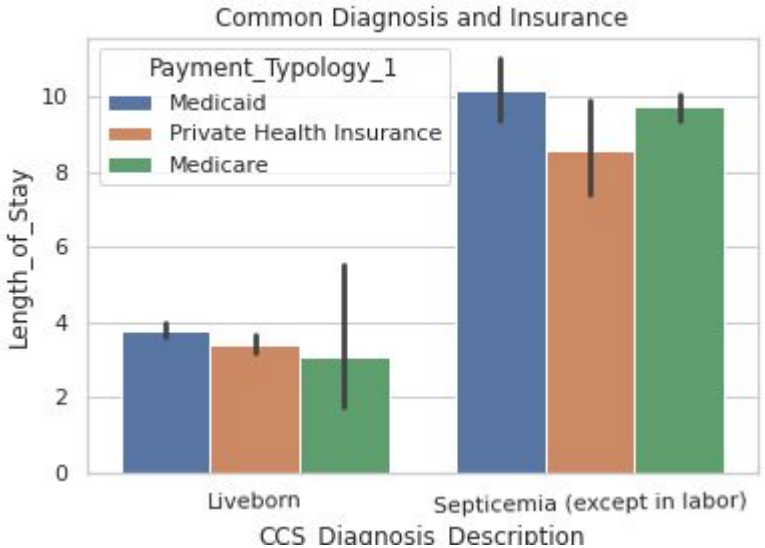
Amongst age groups, the older age groups have a larger variability in their LOS compared to the younger groups. The median LOS in the 50+ age groups is also higher compared to 0-50 years of age.

EDA: Diagnosis Groups



There is significant variability in mean LOS amongst different diagnosis groups. An independent t test on the two commonly occurring Diagnostic groups was statistically significant, yielding a p value < 0.01

Influence of Payer/Insurance and LOS:



Comparison of two commonly occurring diagnosis groups with three common payer types against LOS highlighted some curious facts. There appears to be variability in LOS for the same diagnosis based on the type of Primary Insurance the patient has during admission. This is an important consideration when assessing LOS as a quality metric for a given hospital.

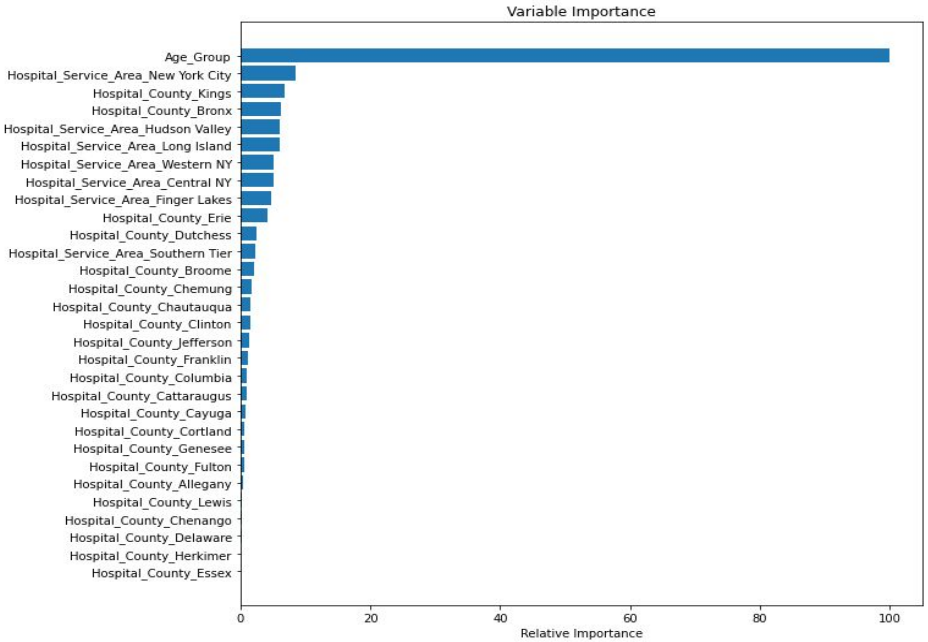
Data Wrangling and PreProcessing:

The target variable LOS was remapped on the basis of ‘below/at mean’ or ‘above mean LOS’. A longer LOS results in higher medical costs for patients and hospitals alike and may also impact insurance reimbursement for hospitals. The class of interest for predicting was therefore the ‘above mean LOS’ which is also the minority class in the dataset.

Categorical features were converted to dummy variable for Logistic Regression and Random Forest Classifier modeling.

Features not useful in predicticting LOS such as Total Charges were dropped. CCS diagnosis and Procedure Codes were utilised.

Feature Importances:



'Age group' comes out as the most significant feature of importance. Following this, interestingly, the 'Hospital service areas' and 'Hospital counties' dominate the feature space. If the Hospital locations are not considered, then the 'Admission-type'(emergency/urgent/newborn), 'Gender', 'Race' followed by 'Diagnosis-type' stand out as important features.

Modeling:

A 5% fraction of the project dataset was grid searched and the best parameters were applied to the ultimate models. The table belows shows a comparison of the different ROC_AUC scores and score of the ‘above_mean’ group at best threshold.

Models	ROC_AUC score	F1 score of ‘above_mean’ LOS at best threshold
Random Forest	0.752	0.62
Logistic Regression	0.751	0.62
CatBoost	0.756	0.63

CatBoost Classifier Classification Report at best threshold:

	precision	recall	f1-score	support
0	0.89	0.80	0.84	17360
1	0.56	0.71	0.63	6250
accuracy			0.78	23610
macro avg	0.72	0.76	0.73	23610
weighted avg	0.80	0.78	0.78	23610

The CatBoost classifier has a better ROC_AUC score. At the best threshold it has a better f1 score for our class of interest compared to the other models.

For class 1 (minority class) the precision was 0.56 and recall was 0.73, with an f1 score of 0.63
Overall accuracy of the model was 0.78

Due to the imbalanced dataset, model performance was assessed by f1 score of the minority class.

Conclusion:

Analysis showed that 'Age Group' and 'Hospital Location' features impacted LOS the most compared to other features. As ‘Hospital County’ and ‘Hospital Service Area’ have more influence on LOS compared to other features such as ‘Diagnosis’, further inquiry is necessary to understand this relation better.

The CatBoost Classifier was the best model for predicting the higher LOS class. This modeling can be adapted for LOS prediction in hospital systems.

Next Steps:

Some factors to incorporate for future modelling:

- Additional information such as Hospital tier type based on type of county and type of service area can be explored.
- Socio economic factors based on geographic location and their influence on Hospital LOS
- Majority payer type based on geographic location

References:

1) Hospital Inpatient Discharges (SPARCS De-Identified) Downloadable File: 2014 | State of New York (ny.gov)

2) (The ABCs of DRGs | ACP Hospitalist)

3) (Medicare and Health Insurance | Office for the Aging (ny.gov))

4)Medicare Guidelines for Inpatient Rehab Coverage (healthline.com)

5) Inpatient or outpatient hospital status affects your costs | Medicare

6) Decreasing the Patient Length of Stay (LOS) to Lower HAIs (centrak.com)

8)https://www.debt.org/medical/hospital-surgery-costs/

9) NYC Health + Hospitals patient care locations - 2011 | NYC Open Data (city of new york.us).

10) https://www.hcup-us.ahrq.gov/toolssoftware/ccs10/ccs10