Hospital Length of Stay Prediction

Introduction

Length of Stay(LOS): Why is this important?

According to debt.org, 60% of all bankruptcies are related to medical expenses. When care was provided for COVID-19 hospitalization without insurance benefits, the charges averaged about $68,261 for the 20 year old age group and $77,323 for patient's 60 years of age (Healthcare Finance news). A longer LOS results in higher average costs, and an increased burden on patient's and their families (American Journal of Managed Care). Hospital LOS is also an important quality metric of hospital efficiency and quality of care.Increased LOS can cause increased risk for healthcare acquired infections and decrease hospital capacity for taking new patients.

Solution:
Goal of the project is to Predict LOS based on multiple variables such as Gender, Race, Diagnosis given at the time of admission. Determination of LOS at patient admission is helpful for hospitals to improve operational workflow efficiency, accurately plan for discharges and decrease readmissions.

Key Beneficiaries:
Hospital CEO's and CFO's, Insurance Industries and patients could all benefit from the findings of this project. By knowing the approximate hospital LOS for a particular diagnosis or procedure, hospital administrators can allocate resources and improve quality of care. Patient's can improve understanding of their potential expenses, and plan for their stay.

The NYSDOH Dataset:

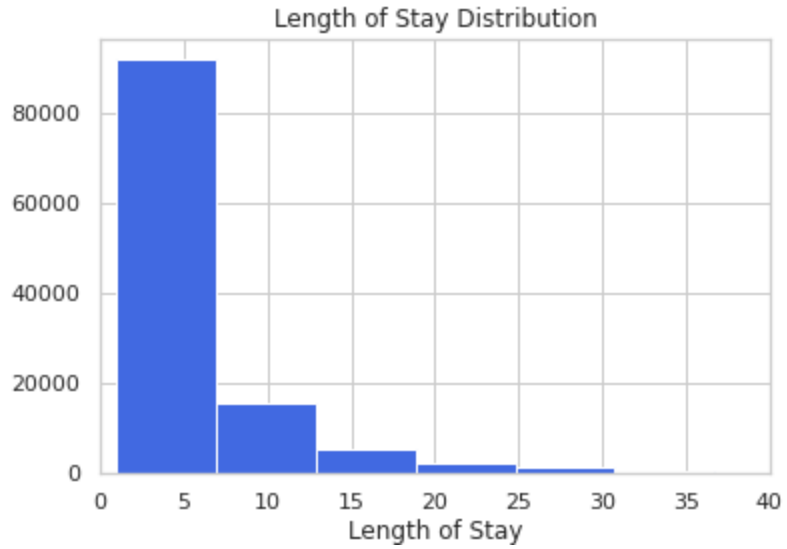
The Statewide Planning and Research Cooperative System (SPARCS) Inpatient De-identified downloadable file contains discharge level detail on patient characteristics, diagnoses, treatments, services, and charges. (2014 data from New York State Department of Health was published 2019). The goal of the project is to predict LOS based on multiple variables such as Gender, Race, Diagnosis given at the time of admission

Data Cleaning:

The original dataset was 911 MB, a 5% random sample from chunks of every 5000 rows of original dataset was taken to use for the capstone project.

EDA Process:

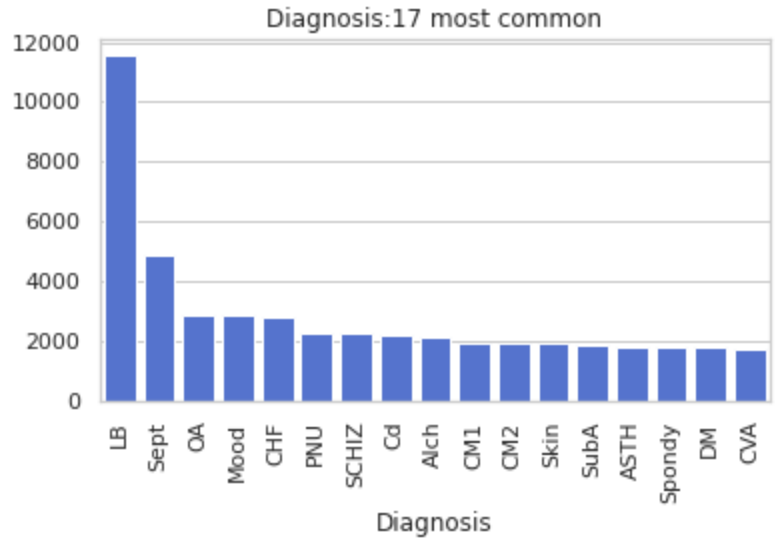
A univariate analysis of LOS was completed. The data were right skewed with the maximum value at 120 + days.



To get a better visualisation of the majority distribution I zoomed in to see where and how the majority of the patient Length of Stay's were distributed. Majority of the distribution is approximately between 1 to 6 days

Figure 1:Frequency distribution of Length of Stay

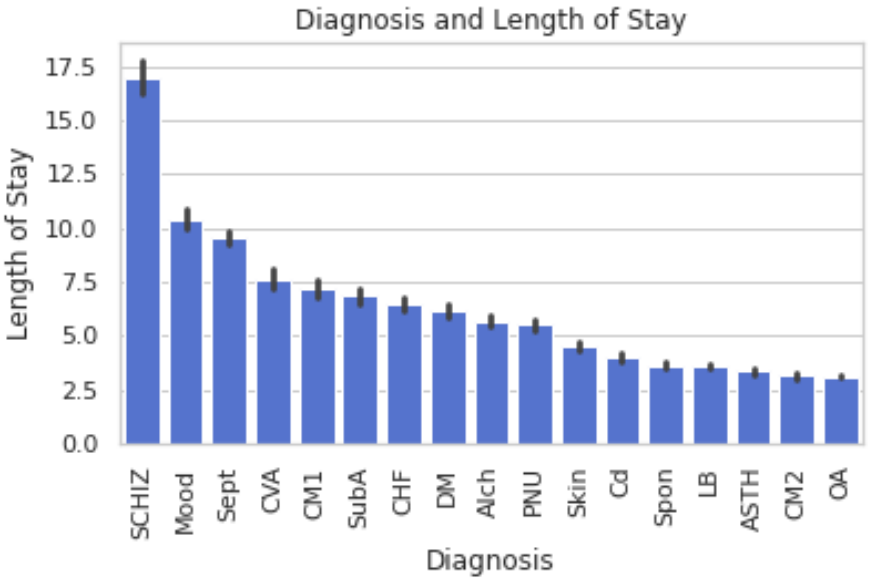
An important consideration was to understand the impact of different Diagnosis and Procedures on LOS. There is a significant variability in LOS amongst 200+ different diagnostic groups. I took a look at the maximal occurring diagnosis groups in hospital admissions.



As seen in Figure 2, the most frequently seen Diagnostic groups are 'Liveborn births', 'Septicemia' and 'Arthritis'. Similar to Diagnostic groups , different Procedures also had significant variability in LOS.

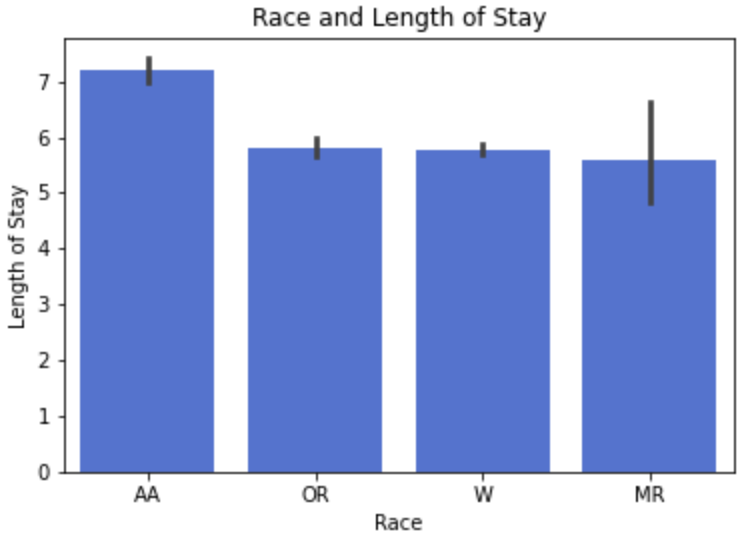
Figure 2: Frequency distribution of the 17 common Diagnosis Groups

An independent t test on the two commonly occurring Diagnostic groups was statistically significant, yielding a p value < 0.01. A statistically significant difference of means was also noted in two commonly occuring Procedure groups.



As seen in Figure 3, Schizophrenia had the highest mean LOS at 17 days followed by Mood disorders at 10 days & Septicemia at ~9 days.

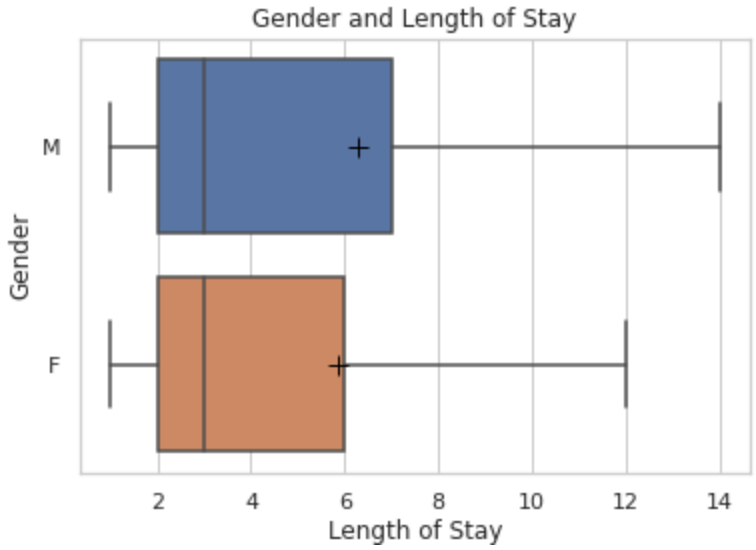
Figure 3: Length of Stay distribution of common diagnosis groups



Amongst different Race groups, the African American population has a higher mean LOS compared to White, Other and Multiracial groups, with Multiracial groups having the most variability.

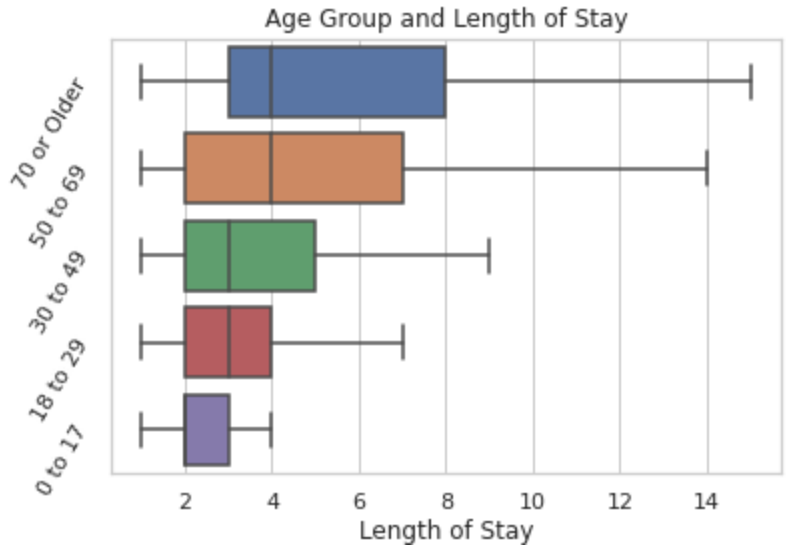
Figure 4: Length of Stay distribution amongst different Race Groups

Besides Race, Diagnosis and Procedure groups, which other factors influence LOS? Let's take a closer look at gender, age, and primary insurance type.



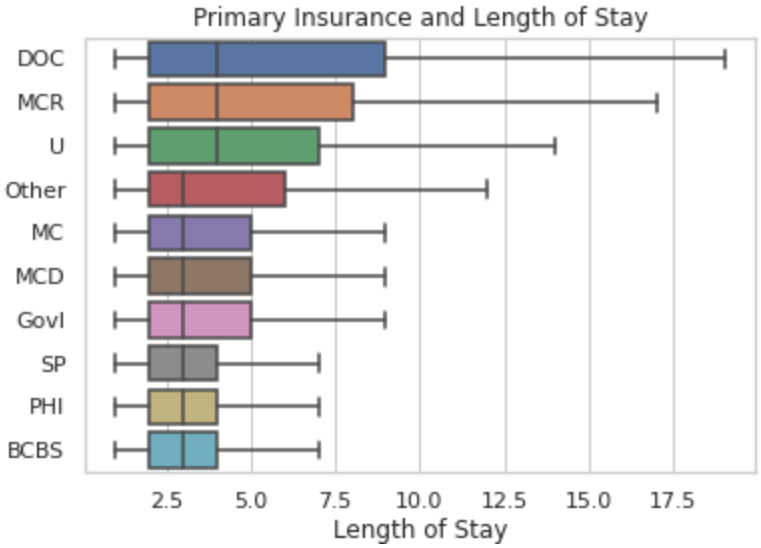
Males had a higher overall LOS compared to females, the difference in means were statistically significant ($p < .01$)

Figure 5: Length of Stay distribution by Gender



Amongst age groups, the overall trend seems to be a higher LOS associated with increased age. One way ANOVA test yields a very low p value (<.001).

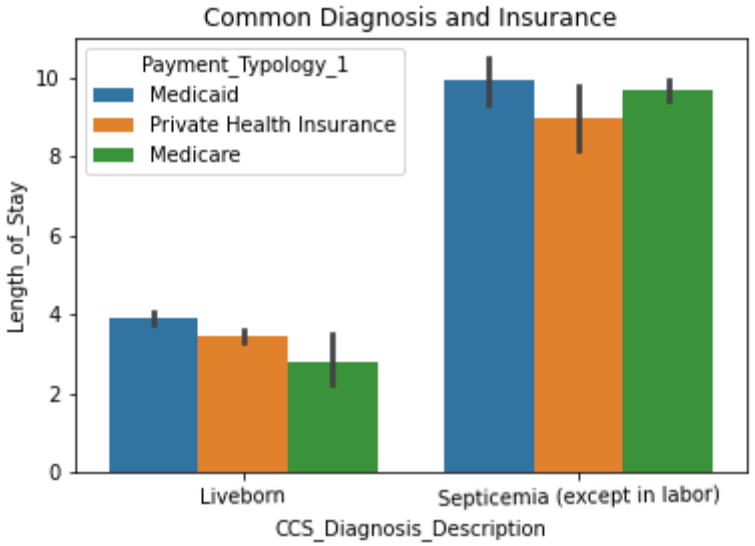
Figure 6: Length of Stay distribution by Age Group



Does the patient's primary insurance, or in other words, ability to pay, influence the LOS? Amongst insurance plans, the Department of Corrections seems to have a higher overall LOS compared to all other insurance groups. The most commonly billed insurance groups however are Medicare, Medicaid and Private Health Insurances.

Figure 7: Length of Stay distribution by Payment Typology 1

After the Department of Corrections, Medicare has a higher mean LOS compared to other insurance groups. A one way ANOVA resulted in a statistically significant p value less than 0.05. A point to note: most people with Medicare are generally 65 years or older. (Medicare and Health Insurance | Office for the Aging (ny.gov)), so age may be a confounding factor here.



Lets evaluate the influence of insurance on LOS for the same diagnosis. Two commonly occurring diagnoses were plotted against LOS while looking at the role of top 3 most frequently utilized payment typology (primary insurance/mode of payment) to see if it made a difference.

Figure 8: Length of Stay variation and interaction between Insurance and Common Diagnosis

This is important to see if LOS is determined by the patient's ability to pay. Interestingly, live births have a lower LOS for Medicare as compared to private insurance, but for Septicemia this is reversed. It's unclear why this might be, but there does appear to be an interaction effect here..These records are reflective of insurances operating in NY state.

<u>EDA takeaways:</u>

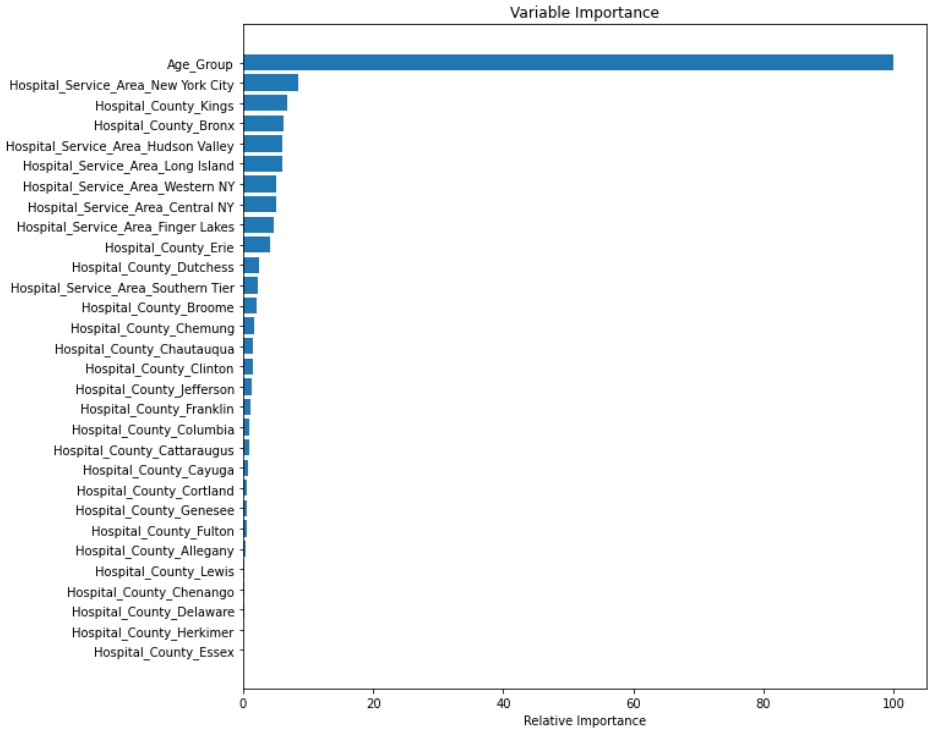
Based on the EDA it is observed that many different variables influence LOS, age group and diagnosis groups being features of interest.

- There is a positive association of increased overall LOS with increased age.

- There is also significant variability amongst LOS in different diagnosis groups,procedure groups and insurance groups.
- Male gender has a higher LOS compared to the female gender
- Majority of LOS numbers lie at or below mean.

Modeling:

The next step was to build an ML classifier to predict LOS. Random Forest, Logistic Regression and CatBoost were grid searched using 5% of the project dataset to pick the best parameters. These best parameters were then applied to the ultimate models and then the models were compared against one another.



Features of importance saw Age group as the most important feature followed by NY city service area and other Hospital Service areas dominating the rest of the feature space. Following location criteria,'emergency room admission criteria', 'race' and 'diagnosis groups' were important. features.

Figure 9: Feature importance from Random Forest Model

The CatBoost Classifier had the best results amongst the models. The metric for comparison was the f1 score for the 'above mean' minority class which was 0.63 with a roc_auc score of 0.76

Why was F1 score of 'above mean' used as a metric?

The dataset is imbalanced with the above mean group being the minority class and also the prediction of interest. For predicting the above mean class, both precision and recall are important, and focussing on only one category will not be advantageous.

Being able to predict which patients would require a longer LOS (i.e above mean LOS) in the hospital is important. A false negative, or predicting incorrectly a shorter LOS when it should have been an above mean LOS can lead to inadequate resources such as lack of hospital beds, nursing staff and medical supplies leading to a decline in quality of care for admitted and incoming patients. A false positive, or predicting incorrectly a longer above mean LOS can result in loss of revenue due to over allocation of resources in certain areas of the hospital and depletion of resources in other essential areas of the hospital (example incorrectly increasing staffing on general medical surgical floor and incorrectly decreasing staffing in pre-op or post anesthesia care unit).

To find a balance the F1 score which is a single metric that combines recall and precision using the harmonic mean was used.

Further Questions and Discussion :

1) Why is geographic location a feature of importance?

Population characteristics, majority payer type (insurance) and socioeconomic factors could all influence how medical services are utilized in certain geographic locations. Having access to this information may help improve our understanding of how these factors influence LOS. Having access to information whereby we can

differentiate hospitals by their designation levels for trauma centers, critical access centers, hospital level 1,2,3,4 etc. and noting if certain hospital levels are more prevalent in certain geographic distributions may be useful.

2) Why was Accuracy Score not used as a metric?

Accuracy Score was not considered a metric of choice due to the imbalance in the prediction classes.

3) APR codes were dropped and CCS codes were used. A little more on why CCS codes were used versus APR-DRG.

- CCS codes:(The Clinical Classifications Software (CCS) for ICD-9-CM is a diagnosis and procedure categorization scheme that can be employed in many types of projects analyzing data on diagnoses and procedures. CCS can be used to identify populations for disease- or procedure-specific studies or to develop statistical reports providing information (such as charges and Length of Stay) about relatively specific conditions. Please note, since October 2015, ICD-10 codes are used which are a revised version of the ICD-9 codes.

- APR-DRGs (The ABCs of DRGs | ACP Hospitalist). In 1983, Medicare adopted the DRG methodology (now known as CMS-DRGs) for hospital inpatient care reimbursement, with the intention of curbing skyrocketing health care costs. All Patient Refined (APR) DRG systems are widely used in the United States for non-Medicare patients. Every year, CMS assigns a “relative weight” to every DRG. The relative weight determines the reimbursement associated with that DRG and reflects the patient's severity of illness and cost of care during hospitalization. A higher relative weight is associated with longer Length of Stay, greater severity of illness, and higher reimbursement). The APR codes may correlate certain diagnosis groups to Length of Stay, but it may not be beneficial in determining the Length of

Stay as it is unclear when these codes are assigned. For this reason, APR categories will not be taken into consideration during the EDA. Total charges, total costs, and total cost to charge ratio are identified after the Length of Stay is complete and thus not useful in predicting Length of Stay. These columns were not considered for the Length of Stay prediction as well.

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