Problem Statement

Recent Covid-19 Pandemic has raised alarms over one of the most overlooked areas to focus: Healthcare Management. While healthcare management has various use cases for using data science, patient length of stay is one critical parameter to observe and predict if one wants to improve the efficiency of the healthcare management in a hospital.

This parameter helps hospitals to identify patients of high LOS risk (patients who will stay longer) at the time of admission. Once identified, patients with high LOS risk can have their treatment plan optimized to minimize LOS and lower the chance of staff/visitor infection. Also, prior knowledge of LOS can aid in logistics such as room and bed allocation planning.

Suppose you have been hired as Data Scientist of HealthMan – a not for profit organization dedicated to manage the functioning of Hospitals in a professional and optimal manner.

The task is to accurately predict the Length of Stay for each patient on a case by case basis so that the Hospitals can use this information for optimal resource allocation and better functioning. The length of stay is divided into 11 different classes ranging from 0-10 days to more than 100 days.

The evaluation metric for this hackathon is 100*Accuracy Score.

Data Findings for any underlying discrepancies:

- No discrepancies in data types of the columns
- In training set we found .03% missing values in Bed Grade and 1.42% missing values in city code patients
- In test set we found .02% missing data in Bed Grade and 1.5% missing in city code patients
- **Data Quality Check :**
    - Proportion is approximately equal for categories of hospital type code between training set and test set.

    - Proportion is approximately equal for categories of City Code Hospital between training set and test set
    - Proportion is approximately equal for categories of Hospital Region Code between training set and test set
    - It is noticed that certain categories of **Available Extra Room** variable of train set are missing from **Available Extra Room** Variable of test set. One way to treat this is to treat the variable as a continuous variable and create bins, or remove the rows having those categories. If we do not treat it , the production model might not perform well.
    - Proportion is approximately equal for categories of Department between training set and test set
    - Proportion is approximately equal for categories of Ward Type between training set and test set
    - Proportion is approximately equal for categories of Ward Facility Code between training set and test set
    - Proportion is approximately equal for categories of Bed Grade between training set and test set
    - Proportion of City Code differs for certain codes between train and test set
    - Proportion is approximately equal for categories of Admission Type between training set and test set
    - Proportion is approximately equal for categories of Severity of Illness between training set and test set
    - Proportion is approximately equal for categories of Age between training set and test set
    - We can see that most people's duration of stay is between 0 days to 2 months . There are very few patients who require more than that.
    - We see the distribution of deposits is the same for both sets. But it appears to be positively skewed , with a long tail towards the right. With outliers the curve appears to be moreover near to normal distribution bell curve
    - Distribution of visitors is skewed positive , but it approximates well for both train and test set. It also suggest presence of outliers,
- There were outliers present in the deposit column, after treating the outliers, there is at least presence of three gaussians, which indicates , data contains deposits of at least 3 different clusters of patients. It can be because of different hospital type as like central , state or private hospital
- There were outliers present in visitors with a patient column, after treating the outliers, there is at least presence of six gaussians, which indicates , data

contains deposits of at least six different clusters of patients. Mostly private hospitals allow 2 visitors per patient + 1 attendant. But it has been seen more in government hospitals , where people are able to dodge the security.

- For Missing Values have imputed data :
    - Categorical Variable : Mode Imputation
    - Numerical Variable : -999

Insights

- Hospital Type ‘a’ has the highest probabilities for each category of stay

```
pd.crosstab(train['Hospital_type_code'],train['Stay'],normalize=True)
```

Stay	0-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80	81-90	91-100	More than 100 Days
Hospital_type_code											
a	0.033159	0.125005	0.126510	0.075287	0.012910	0.045903	0.002902	0.011688	0.005935	0.003021	0.008080
b	0.016157	0.043752	0.061363	0.038893	0.012225	0.021697	0.002880	0.008190	0.003222	0.002459	0.005672
c	0.011268	0.032402	0.039336	0.025820	0.005879	0.016255	0.001350	0.005059	0.002277	0.001379	0.003206
d	0.004126	0.013890	0.017357	0.011638	0.002387	0.008520	0.000584	0.002390	0.001322	0.000565	0.001250
e	0.005668	0.020865	0.020488	0.012750	0.002399	0.008894	0.000619	0.002588	0.001240	0.000685	0.001592
f	0.003213	0.007323	0.005957	0.006573	0.000641	0.006155	0.000138	0.001514	0.000798	0.000433	0.000867
g	0.000534	0.002142	0.003737	0.002258	0.000437	0.002547	0.000144	0.000773	0.000399	0.000141	0.000320

- We can see almost all the city code hospital have max stay of between 21-30 days apart from city code hospital =7

City Code Hospital	Max Stay Duration	Max Prob
1	21-30	.051
2	21-30	.047
3	21-30	.027
4	21-30	.011
5	21-30	.026
6	21-30	.045
7	11-20	.037
9	21-30	.019
10	21-30	.0041
11	21-30	.0013
13	21-30	.003571

- We can see the hospital code also has the highest probability for stay of between 21-30 days

Hospital Code Region	Max Stay Duration	Prob
X	21-30	.12
Y	21-30	.099
Z	21-30	.054

- There is high possibility of 4 rooms available for stay duration of between 21-30 days
- For stay duration of 21-30 , sorted in descending on the basis of their proportion , departments are as follows :
    - Gynecology
    - Anesthesia
    - Radiotherapy
    - TB & Chest disease
    - Surgery

Department	0
gynecology	0.214742
anesthesia	0.028106
radiotherapy	0.022808
TB & Chest disease	0.008221
surgery	0.000870

    - Ward Type : R,Q,P,T,U , in descending order of their proportion have higher proportion for stay duration of 21 to 30 days, only Ward Type : S has higher proportion for stay duration of 11-20 days
    - Ward Facility Code in descending order of their proportion : F,E,D,B,A have higher proportion for stay duration of 21-30 days , whereas only Ward Facility Code : C has higher proportion for stay duration of 11-20 days.
    - Bed Grade in descending order of their proportion for stay duration between 21-30 are : 2 & 1, and for stay duration between 11-20 , bed grade are : 3,4

- Top 5 city code of patients are 8,2,1,7,5, and as per the same order their stay is between 21-30 days

City_Code_Patient	0
8.0	0.109272
2.0	0.037317
1.0	0.024237
7.0	0.019687
5.0	0.015692

- For Admission Type : Trauma and Urgent, patient generally stays for duration of 21-30, whereas for Admission Type: Emergency , generally stay of duration is 11-20 days
- For Moderate and Extreme severity of illness general stay duration is 21-30 days , whereas for Minor Severity of Illness stay duration is 11-20 days
- For all Severity, general admission type is Trauma
- For age group of between 0 to 20 general stay duration is between 11-20 days , whereas for, age group above 20, duration of stay generally is between 21-30

- From below we can see that apart from age groups 81-90 , the rest age group has a general duration of stay of 21-30 days and has max mean deposit during that duration. Only for the age group of 81-90, general duration of stay is 61-70 days with max mean deposit.

Age Group	General Duration of Stay	Mean Deposit
91-100	21-30	5237
81-90	61-70	5182
71-80	21-30	5147
61-70	21-30	5109
0-10	21-30	5077
11-20	21-30	5074
51-60	21-30	5015
21-30	21-30	5002
31-40	21-30	4953
41-50	21-30	4928

- We have already seen for all the categories of severity of illness , general duration of stay is between 21-30 days, but mean amount deposited is :

Severity_of_Illness	
Minor	5124.170142
Moderate	5003.759395
Extreme	4891.296768

- We have already seen the general stay duration for all the admission type is 21-30 days, their average amount deposited is :

	0
Type_of_Admission	
Trauma	5086.610017
Emergency	4939.604785
Urgent	4915.376399

- Top 5 City Code of Patient with maximum mean deposit are : 30,31,29,11,32

City_Code_Patient	0
30.0	7243.5
31.0	6598.5
29.0	6302.0
11.0	5645.0
32.0	5541.0

- General stay duration is of 21-30 days for each bed grade category. Their average deposit by bed grade :

Bed_Grade	
-999.0	5226.000000
4.0	5142.025128
3.0	5099.145388
2.0	4965.118902
1.0	4773.433599

**-999 indicates the missing bed grade

- For ward facility type : A,B,C,D,F for their general stay duration of 21-30 days the average amount deposit . Ward Type E deposit is for a stay duration of 11-20 days.

Ward_Facility_Code	
A	5190.742442
F	5019.202442
B	5018.982733
D	5016.768362
E	4953.044659
C	4942.884899

-

Ward Type	Average Deposit	Stay
S	5148	21-30
R	5093	21-30
P	4954	61-70

Q	4890	21-30
T	4836	11-20
U	4527	21-30

-

Department	Average Deposit	Stay Duration
anesthesia	5407	21-30
TB & Chest disease	5330	41-50
surgery	5136	21-30
radiotherapy	5071	21-30
gynecology	4946	21-30

-