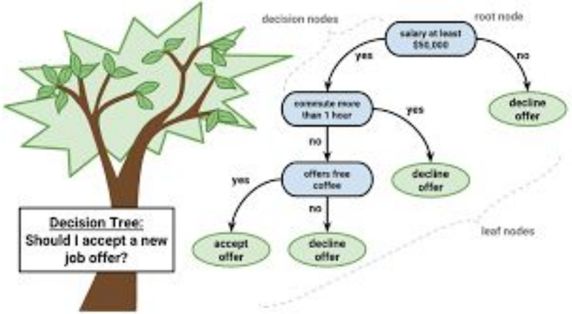
UMBC

Use COVID hospital treatment data to Predicting length of stay

DATA 603
Alex Zhou

Introduction

The COVID-19 pandemic has placed an unprecedented strain on health systems, with rapidly increasing demand for healthcare in hospitals and intensive care units (ICUs) worldwide. As the pandemic escalates, determining the resulting needs for healthcare resources (beds, staff, equipment) has become a key priority for many countries. Projecting future demand requires estimates of how long patients with COVID-19 need different levels of hospital care.



Dataset Description

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 miminize 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.

```
root
 |-- case_id: string (nullable = true)
 |-- Hospital: string (nullable = true)
 |-- Hospital_type: string (nullable = true)
 |-- Hospital_city: string (nullable = true)
 |-- Hospital_region: string (nullable = true)
 |-- Available_Extra_Rooms_in_Hospital: string (nullable = true)
 |-- Department: string (nullable = true)
 |-- Ward_Type: string (nullable = true)
 |-- Ward_Facility: string (nullable = true)
 |-- Bed_Grade: string (nullable = true)
 |-- patientid: string (nullable = true)
 |-- City_Code_Patient: string (nullable = true)
 |-- Type of Admission: string (nullable = true)
 |-- Illness_Severity: string (nullable = true)
 |-- Patient_Visitors: string (nullable = true)
 |-- Age: string (nullable = true)
 |-- Admission_Deposit: string (nullable = true)
 |-- Stay_Days: string (nullable = true)
```

Data observation

stay_days	count
21-30	87491
11-20	78139
31-40	55159
51-60	35018
0-10	23604
41-50	11743
71-80	10254
More than 100 Days	6683
81-90	4838
91-100	2765
61-70	2744

illness_severity	count
Moderate	175843
Minor	85872
Extreme	56723

age	count
41-50	63749
31-40	63639
51-60	48514
21-30	40843
71-80	35792
61-70	33687
11-20	16768
81-90	7890
0-10	6254
91-100	1302

department	count
gynecology	249486
anesthesia	29649
radiotherapy	28516
TB & Chest disease	9586
surgery	1201

Counts of rows/samples: 318438
Counts of columns/features: 18

Filtering/Method/Parameter Setting

First,Split the data 80/20,train/test, In order to train ML models in Spark later, I use the VectorAssembler() to combine a given list of columns into a single vector column. Next, we standardize the features by StandardScaler(), After the preprocessing step, I fit the PCA() (Principal Component Analysis) model to reduce the dimensionality of large data sets.

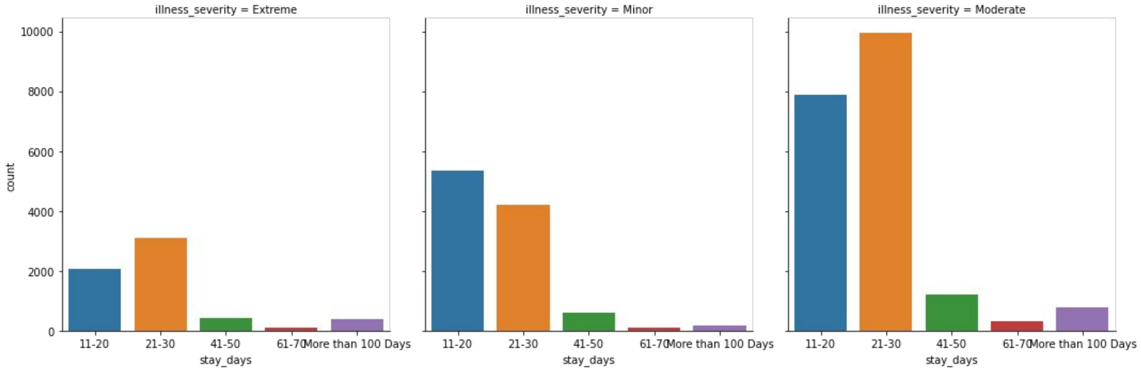
Parameters of DecisionTreeClassifier
Impurity - Gini
maxBins - 24,32
minInfoGain - 0.0, 0.2
maxDepth - 5,10
Cross-validator 10 fold

Result

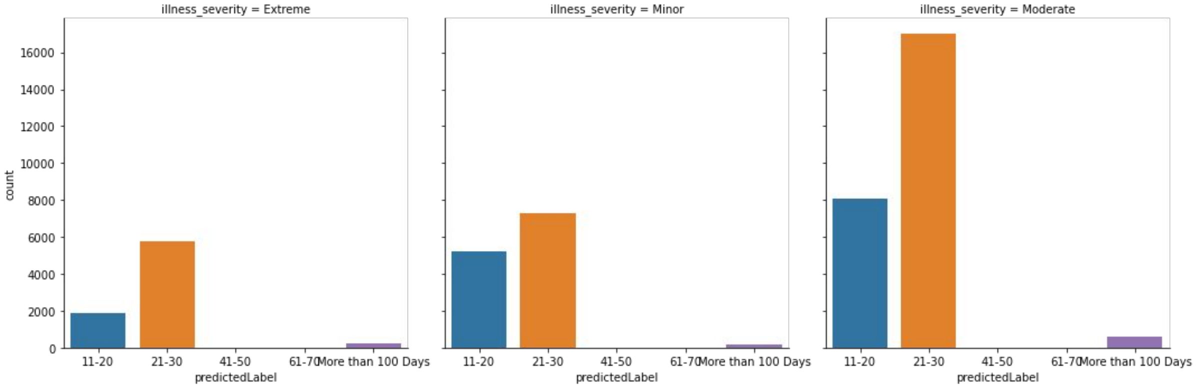
case_id	illness_severity	type_of_admission	department	stay_days	predictedLabel
1	Extreme	Emergency	radiotherapy	0-10	21-30
10001	Minor	Trauma	gynecology	11-20	21-30
100011	Minor	Trauma	gynecology	21-30	21-30
100013	Minor	Trauma	gynecology	31-40	51-60
100014	Minor	Urgent	gynecology	81-90	51-60
100018	Moderate	Emergency	gynecology	41-50	51-60
10002	Minor	Trauma	gynecology	21-30	11-20
100022	Moderate	Trauma	gynecology	11-20	21-30
100029	Moderate	Trauma	gynecology	21-30	21-30
10003	Minor	Trauma	gynecology	21-30	21-30
100031	Moderate	Emergency	gynecology	0-10	11-20
100033	Extreme	Trauma	gynecology	51-60	21-30
100034	Extreme	Trauma	gynecology	91-100	51-60
100038	Extreme	Trauma	gynecology	41-50	21-30
100040	Moderate	Trauma	gynecology	21-30	21-30
100045	Extreme	Trauma	gynecology	11-20	21-30
100047	Extreme	Trauma	anesthesia	51-60	31-40
100049	Moderate	Trauma	radiotherapy	11-20	21-30
100052	Moderate	Trauma	gynecology	21-30	21-30
100054	Moderate	Trauma	gynecology	21-30	21-30

```
Model name: DecisionTreeClassifier
        Accuracy = 0.392973
        F1 = 0.350573
        Test Error = 0.607027
        True Positive Rate By Label = 0.693876
        False Positive Rate By Label = 0.399571
        Precision By Label = 0.398808
        Recall By Label = 0.693876
        FMeasure By Label = 0.506502
        Weighted Recall = 0.392973
        Weighted Precision = 0.362352
        Weighted True Positive Rate = 0.392973
        Weighted FMeasure = 0.350573
        Log Loss = 1.77738
        Hamming Loss = 0.607027
```

Result



illness_severity = Extreme
illness_severity = Minor
illness_severity = Moderate
count
10000
8000
6000
4000
2000
0
11-20
21-30
41-50
61-70
More than 100 Days
stay_days



illness_severity = Extreme
illness_severity = Minor
illness_severity = Moderate
count
16000
14000
12000
10000
8000
6000
4000
2000
0
11-20
21-30
41-50
61-70
More than 100 Days
predictedLabel

TO DO/Solution

Since my data is more features and less samples. This situation greatly increases the difficulty of training. Despite my attempts to reduce the dimensionality of the data and more cross-validations, I improved the accuracy from 20% to 40%. After my research, maybe Pruning or post-pruning is a good way to improve.