Predicting Length of Stay and Total Charges for Lung Cancer and Mentally Ill Patients

Harika Polaki

Table of Contents

1. Introduction

Cancer is a deadly disease caused when cells in our body grow out of control. When this happens in the human lungs, we call it as Lung Cancer. Starting from the lung's cancer may slowly spread to other organs in the body. The primary reason for lung cancer is tobacco smoking. About 10% of cases occur in people who don't smoke. The reason can be genetic factors or getting exposed to harmful chemicals. So, avoiding air pollution and smoking is primary prevention. The process of spreading cancer from one organ to another is known as Metastases. Majorly lung cancer is categorized into two parts, i.e., Small cell lung cancer (SCLC) and Non-small cell lung cancer (NSCLC). The symptoms of lung cancer are cough, chest pain, shortness of breath, blood coughs, weight loss. Usually, lung cancer can be treated in many ways like Surgery, Chemotherapy, Radiation Therapy. Lung cancer patients who have undergone resections are expected to have 58% more odds of postoperative mortality. (American Cancer Society, n.d.)
Studies prove that cancer patients' developmental conditions, such as mental disorders, are commonly known as Mental Illness, Anxiety, and Lack of Confidence. This is observed in patients after undergoing chemo sessions that are held after Staging. So, it is a well-observed thing that there is a link between Lung cancer and Severe Mental Illness (SMI). The aggregation of lung cancer and SMI leads to a greater chance of mortality.

1.1 Background and Rationale

Cancer surveillance data from the CDC and NCI says that there are more than 28 million cancer cases alone in the US. There are about 2,28,820 lung cancer cases by 2020 and about 135,720 deaths caused by lung cancer. Lung cancer deaths have become by far the leading reason for cancer-related deaths among both men and women, making up almost 25% of all cancer deaths. In general, about 13% of deaths are because of SCLC, and 84% are due to NSCLC. On a positive note, the lung cancer risk is reducing in the people who quit smoking. Thus, we believe that there's a need to find a correlation between lung cancer and SMI and the factors affecting the death rate and mortality or immortality chances. (JAMA NETWORK, 2013)

1.2 Project Objective

Upon completing the project, the public will use our analysis to find the correlation between lung cancer and SMI, find about the cost involved with the treatment. As discussed earlier, that lung cancer, in combination with mental illness, is a deadly disease. Our primary objective of this study is to find the correlation between SMI and lung cancer deaths. Finding which factors in SMI have more effect, are they more effective individually or collectively. Through this, we can find the death rate of lung cancer patients with a history of SMI. Eventually, we would like to see the cost curve of the treatment and length of stay for patients affected by these diseases.

2 Data Acquisition

2.1 Overview

For this project, we acquired our datasets from the Healthcare Cost and Utilization Project (HCUP). HCUP is a family of health databases and related software tools used for development, sponsored by the Agency for Healthcare Research and Quality (AHRQ). HCUP has the most extensive collection of health data starting from the year 1988. HCUP databases include The National Inpatient Sample (NIS), The Nationwide Ambulatory Surgery sample (NASS), The State Inpatient Database (SID), and many more. AHRQ creates the HCUP databases through federal-state partnership, and NRD is a unique database created to support various analyses regardless of the expected payer for the hospital stay. Our datasets are from the NIS database, which derives its data from billing data submitted by hospitals statewide across the US. Each state has a SID which reports all its data to NIS. The inpatient records contain all the clinical data. (Services, n.d.)

In this project, we are mainly focusing on NIS 2016 and NIS 2017 data. NIS 2016 and NIS 2017 are ASCII formatted data files on HCUP, where we can find data related to patient diagnosis and procedure codes in ICD-10-CM/PCS format. ICD-10- CM is the updated version of ICD-9-CM, which stands for International Classification of Disease code for Clinical Modifications. It deals with diagnosis coding on inpatient and out-patient data. ICD-10- PCS stands for International Classification of Disease code for Procedure Classification System.

For 2016, there are three data-related files,

-NIS 2016 Core- This data file contains samples of hospital discharge records of various states.

-NIS 2016 Severity- It works in conjunction with the inpatient core file.

-NIS 2016 Hospital- This contains weights and variance estimation data elements.

2.2 Data description

Finding a suitable dataset for performing machine learning algorithms was the first challenging task. In the current state of the healthcare data world, most of the datasets can be unstructured. Most patient-level data might not be available publicly due to privacy reasons. Our sponsors, Allwyn Corporation, provided us three datasets (core, hospital, and severity). When we unzipped these three files, it got extracted too.ASC format. We happened to download 3 STATA load files (.Do format) from the HCUP website, consisting of the attributes. We merged all the six files and made them into 3 data files. The core file consists of data related to patients. Each patient may have multiple records, and the Hospital file has documents of hospitals and the services provided by them to the patients. We merged both these files on the HOSP_NIS attribute and made it into a single file. The Severity file combined with the output file on the KEY_NIS attribute. Finally, we were able to merge 3 data files into 1 STATA(.dta) file. By using big data processing and extraction techniques such as python, we could extract 7.1 million records. Later we exported this STATA file to a CSV file. After shipping, we loaded the file into python using the Pandas library. Field Description

As mentioned earlier, we got our data from the HCUP website. We have NIS data related to the years 2016 and 2017 in which we have three separate files for each year know as Core, Hospital, and Severity. We combined three files each year into one file and performed the analysis. The file contains fields related to patients like admission date, age, sex, race, residence. It even consists of ICD-10 codes, I10_DX1 to I10_DX40, which says about what kinds of diseases the patient diagnosed with, contains procedure codes from I10_PR1 to I10_PR25. Fields related to total charges, Length of stay, discharge information, transfer are also present from which we can fetch the data to find information related to the cost of treatment. We found out that we have a few fields with null values, so we performed some preprocessing to handle the missing values. Though we had all the information we needed, we had to research the elixhauser codes to know about them and categorize them. As we know that clinicians will use ICD-10-PCS for predicting surgical outcomes, we will need this data to pair with ICD-10-CM diagnosis data because of the lack of disease-specific information in the new procedural codes.

ICD-10-CM (String) – I10_DX 1 to I10_DX40- Diagnoses, principal and secondary, with the external cause of morbidity codes at the end of the array.

ICD-10-PCS (String)- Procedures, principal, and secondary codes.

HOSP_NIS (INT)- NIS hospital file linked to hospital file.

DRG (String)- Discharge related groups.

HOSP_DIVISION (VARCHAR) – Division to which the hospital belongs.

MDC (String)- Major Diagnostic Category.

TOTCHG (INT)- Total charges for the entire stay period in the hospital.

LOS (INT)- Length of stay of the patient diagnosed.

KEY_NIS (INT) – Unique record number for file beginning in 2012 links the Core File to other discharge-level NIS files.

AGE_NEONATE (INT)- Neonatal Age (first 28 days after birth) indicator.

FEMALE (BINARY)- 1 if the patient is female, 0 if the patient is male.

2.3 Data Quality Assessment

- **Completeness:** Completeness is a problematic term to discuss in terms of this project. The current dataset which we have is complete in providing us the information we need. But we have a few missing terms on which we performed preprocessing. We are using various diagnosis and procedure codes; collected this information from multiple hospitals across the United States.

- **Accuracy:** There are few fields in the data where they are null for almost all the patients, so we ignore these fields. So, we needn't consider those codes for data modeling as they create noise and create an issue while training the model.

- **Overall Quality:** Overall, the dataset obtained from our data sources is of good quality and provides all the information we need.

- **Atomicity:** A data item is atomic if all the representations of that field across the data set to match.

- **Overall Quality:** The degree to which the sample dataset accurately represents the whole population to be measured.

3 Analytics and Algorithms

3.1 Exploratory Analysis

As discussed earlier, there are so many null columns in the dataset. So, we ignored these null columns, and we replaced them with median values for values that were missing. Handling outliers has been the most significant task in the analysis part. We implemented a few visualizations to understand the data. In the dataset, there are many ICD-10-CM/PCS codes. We had to filter lung cancer and mental illness codes to perform analysis.

ICD-10-PCS Procedure codes for Lobectomy:

0BTC0ZZ
0BTC4ZZ
0BTD0ZZ
0BTD4ZZ
0BTF0ZZ
0BTF4ZZ
0BTG0ZZ
0BTG4ZZ
0BTH0ZZ
0BTH4ZZ
0BTJ0ZZ
0BTJ4Z

ICD-10-PCS Procedure Codes for Mental, Behavioral and Neurodevelopment disorders:

F40-F48- Anxiety, dissociative, and other non-psychotic mental disorders.
F60-F69- Disorders of adult personality and behavior.
F50-F59- Behavioral syndromes with psychological syndromes.
F80-F89- Pervasive and specific developmental disorders.
F30-F39- Mood disorders.
F20-F29- Schizophrenia, Schizotypal, and mood psychotic disorders.
F90-F98- Behavioral and emotional disorders.

F10-F19- Mental and behavioral disorders.
F01-F09- Mental disorders.

- **Average Total charges and Length of stay for different location status**

Location status	Total charges	Length of stay
Rural	67964	6
Urban non-teaching	81990	6
Urban teaching	79348	5

Table 1

From the above analysis, we can understand that cost of treatment and length of stay are more in urban non-teaching than in other areas. We can also conclude that hospitals in rural and urban non-teaching have the highest Length of stay than urban teaching zone where both cost and length of stay are less comparatively. From the above analysis, we got Pearson's correlation as 0.568. This means there is a positive correlation between LOS and Total Charges.

- **Percentage of Male and Female lung cancer patients**

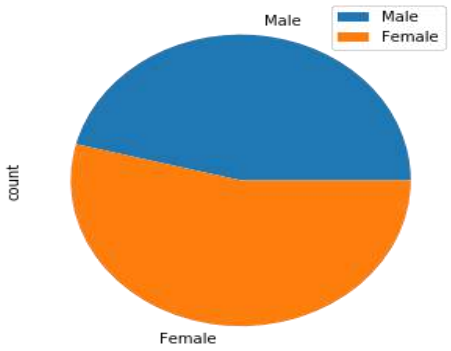

Fig 1

Gender	**Count**
Male	**46**
Female	**54**

Table 2

We see that lung cancer is more in females with a percentage of 54%.

- **Lung Cancer Patients in different regions**

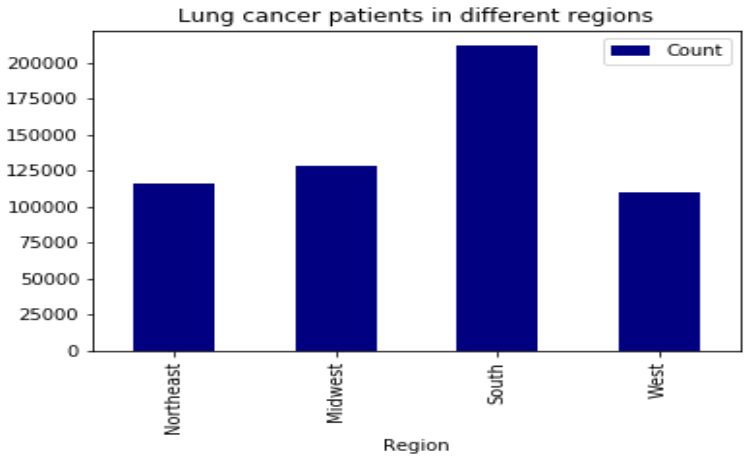

Fig 2

From the above analysis, we can conclude that the south region has the highest number of cases than others.

- **Grouping Mental illness Dx codes after filtering Lobectomy procedure codes**

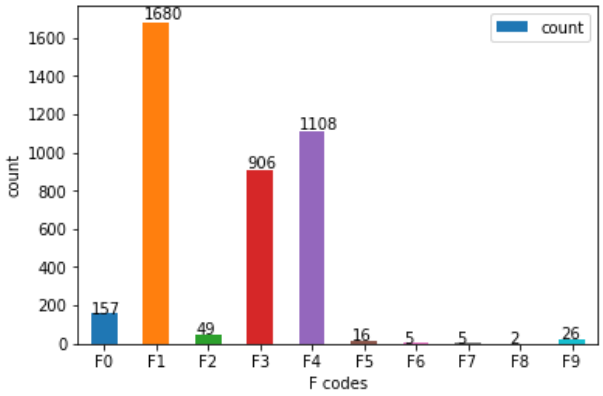

Fig 3

Grouping Mental illness Dx codes after filtering Lobectomy procedure codes, we found that groups of patients with Mental disorders due to substance use have the most lung cancer than other groups.

- **Number of patient's vs. Different types of Lobectomy**

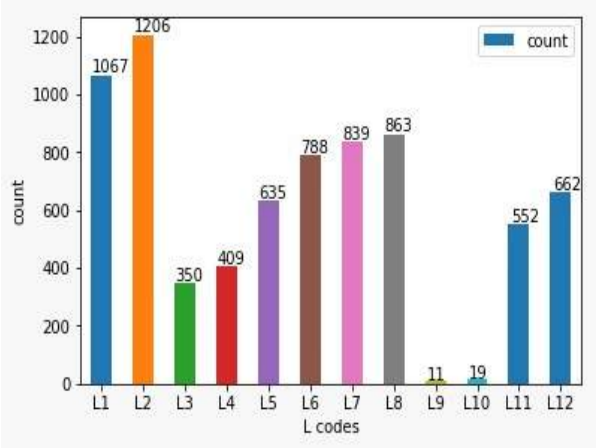

Fig 4

From the above visualization, we can observe that many patients are suffering from upper lung lobe, right lobectomy with endoscopic (L2- 0BTC4ZZ) followed by Open approach(L1-0BTC0ZZ).

4 Findings

4.1 Problem Statement 1- Finding a correlation between Mental Illness and Lung Cancer-related Deaths.

- **The co-relation between Mental Illness and Lung Cancer-related deaths**

Heat Map- Co-relation Plot

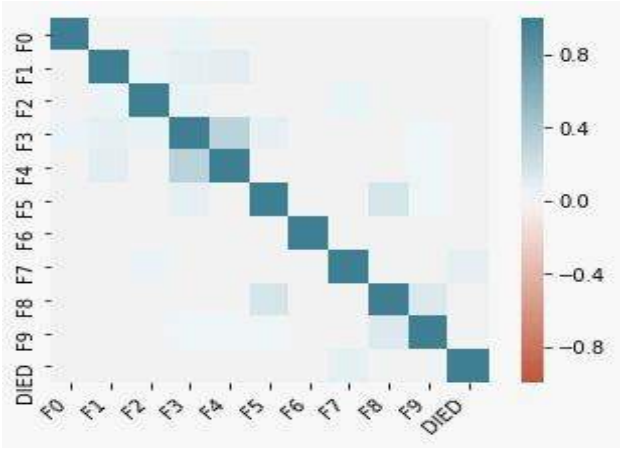

Fig 5

We cannot see any strong correlation between mental illness and lung cancer with deaths from the above heat map. So, we moved to Scatterplot heat map to interpret results in a better way.

Heat Map- Scatter Plot Representation

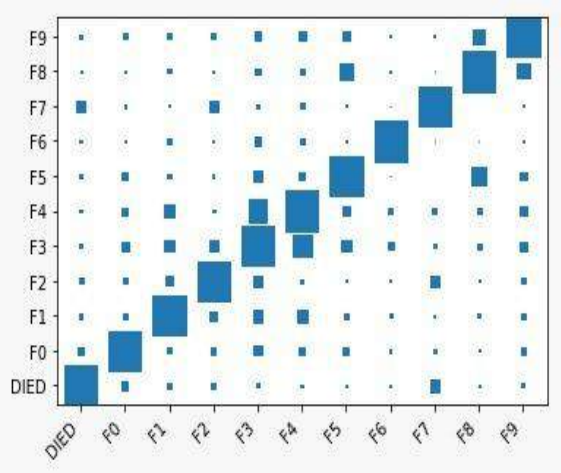

Fig 6

From the above scatter plot representation, we can observe a slight likely correlation at F0 and F7, but still, we cannot conclude it is positive or negative. So, we planned to implement a Coefficient Heat map representation.

Heat Map- Coefficient Representation

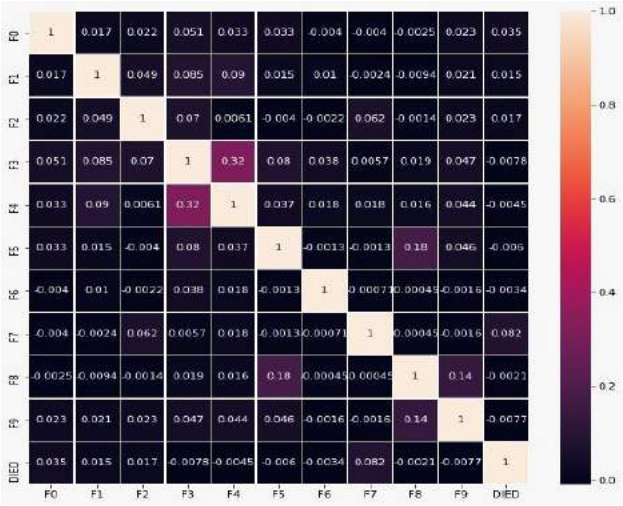

Fig 7

From the above Coefficient representation, we can observe that all the coefficients are close to zero, but there is a slight likely correlation at F7 to Death.

On performing the above correlation plots, we observed a slight correlation with F0 and F7 with mental illness and lung cancer deaths. So, we moved forwarded on implementing Chi-Square Tests on individual mental illness categories with fatalities.

The Chi-square test is a statistical test used to determine whether the output variable is dependent or independent of the input variable. We already filtered the mental illness codes, which are also diagnosed with lung cancer. So, we perform a chi-square test on individual F codes to determine their dependency on Death. After completing the Chi-square test, we will be estimating our results based on Pearson's Chi-Square coefficient, p-value, and Cramer's coefficient. If the p-value is less than 0.05, we can reject the null hypothesis, and if the p-value is more significant than 0.05, we can accept the null hypothesis. Cramer's coefficient describes the strength of association between the variables. If the value is between 0 and 0.1, it means there is a little association. If it is between 0.1 and 0.3, there is a low association; if it is between 0.3 and 0.5, there is a moderate association. If the value is more significant than 0.5, it means there is a high association. (Brownlee, 2018)

Cramer's Coefficient Table:

Describing Strength of Association

Characterizations	
>.5	high association
.3 to .5	moderate association
.1 to .3	low association
0 to .1	little if any association

Fig 8

Chi-Square Test Results:

- **Test Results for F0- Mental Disorders due to psychological conditions**

```
test_results_f0
```

	Chi-square test	results
0	Pearson Chi-square (1.0) =	10.5661
1	p-value =	0.0012
2	Cramer's phi =	0.0435

Fig 9

P-value < 0.05 so we can reject the null hypothesis that mental disorders due to psychological conditions, F0 and Death are related to each other. Cramer's coefficient is less than 0.3, which means there is a little association.

- **Test Results for F1- Mental and behavioral disorders due to psychoactive substance use**

```
test_results_f1
```

	Chi-square test	results
0	Pearson Chi-square (1.0) =	0.0902
1	p-value =	0.7639
2	Cramer's phi =	0.0040

Fig 10

P-value > 0.05, so we cannot reject the null hypothesis that F1 and Death are independent variables.

- **Test Results for F2- Schizophrenia, schizotypal, delusional, and other non-mood psychotic disorders**

```
test_results_f2
```

	Chi-square test	results
0	Pearson Chi-square (1.0) =	1.6536
1	p-value =	0.1985
2	Cramer's phi =	0.0172

Fig 11

P-value > 0.05, so we cannot reject null hypothesis that means F2 and Death are independent variables

- **Test Results for F3- Mood [affective] disorders**

```
test_results_f3
```

	Chi-square test	results
0	Pearson Chi-square (1.0) =	1.4203
1	p-value =	0.2333
2	Cramer's phi =	0.0160

Fig 12

P-value > 0.05, so we cannot reject the null hypothesis that means F3 and Death are independent variables.

- **Test Results for F4- Anxiety, dissociative, stress-related, somatoform, and other non-psychotic mental disorders**

```
test_results_f4
```

	Chi-square test	results
0	Pearson Chi-square (1.0) =	1.3387
1	p-value =	0.2473
2	Cramer's phi =	0.0155

Fig 13

P-value > 0.05, so we cannot reject the null hypothesis that means F4 and Death are independent variables.

- **Test Results for F5- Behavioral syndromes associated with physiological disturbances& physical factors**

```
test_results_f5
```

	Chi-square test	results
0	Pearson Chi-square (1.0) =	0.3620
1	p-value =	0.5474
2	Cramer's phi =	0.0081

Fig 14

P-value > 0.05, so we cannot reject the null hypothesis that means F5 and Death are independent variables.

- **Test Results for F6- Disorders of adult personality and behavior**

```
test_results_f6
```

	Chi-square test	results
0	Pearson Chi-square (1.0) =	0.1149
1	p-value =	0.7347
2	Cramer's phi =	0.0045

Fig 15

P-value > 0.05, so we cannot reject the null hypothesis that means F6 and Death are independent variables.

- **Test Results for F7- Mental Retardation**

```
test_results_f7
```

	Chi-square test	results
0	Pearson Chi-square (1.0) =	24.1331
1	p-value =	0.0000
2	Cramer's phi =	0.0658

Fig 16

P-value < 0.05 so we can reject the null hypothesis that means that F7 and Death are related to each other. Cramer's coefficient is less than 0.3, which means there is a little association.

- **Test Results for F8- Pervasive and specific developmental disorders**

```
test_results_f8
```

	Chi-square test	results
0	Pearson Chi-square (1.0) =	0.0820
1	p-value =	0.7746
2	Cramer's phi =	0.0038

Fig 17

P-value > 0.05, so we cannot reject the null hypothesis that means F8 and Death are independent variables.

- **Test Results for F9- Behavioral and emotional disorders with onset usually occurring in childhood and adolescence**

```
test_results_f9
```

	Chi-square test	results
0	Pearson Chi-square (1.0) =	0.8102
1	p-value =	0.3680
2	Cramer's phi =	0.0120

Fig 18

P-value > 0.05, so we cannot reject the null hypothesis that F9 and Death are independent variables.

From these results, we can conclude that F0(Mental Disorders due to psychological conditions) and F7(Mental Retardation) show a positive correlation with Death.

4.2 Problem Statement 2- Predicting Length of Stay for Patients with SMI and Lung Cancer

The cost of hospital stays in the USA health system costs at least $375 billion per year. Recent Medicare legislation standardizes payments for procedures performed regardless of the number of days a patient spends in the hospital. This demands knowing patients with high Length of stay (LOS) at the time of admission. If patients get to know about LOS at the time of admission, they can optimize the treatment plan to minimize their LOS, and this also helps hospitals to plan rooms and beds for patients.

The goal here is to create the best model to predict the length-of-stay for each patient at the time of admission. The model's inputs include all the features or details about patient like Age, sex, marital status, and diagnosis category. So, we have to prepare the input dataset in such a way that it yields better results when an algorithm is applied.

Predicting Length-Of-Stay (LOS):

This project's primary aim is to develop a model that is better at predicting LOS than the industry standards of median and average LOS. To measure the model's performance, the expected model is compared against the median and average LOS using root means square error (RMSE). Low RMSE means greater accuracy of the model. The ultimate goal is to develop a model with low RMSE models than average or median LOS models. There are multiple regression models available to predict LOS, and R2 (R-Squared), which is the measure of goodness of the model, will be used to predict the model. The best possible R2 is 1.0, and a negative value means it is worse than a constant model. (codes)

Firstly, we combined both NIS 2016 and NIS 2017 data, and we tried understanding our data, as we aim to predict LOS. Below we can see that there are 5581 records for LOS, min LOS of 1 day, max LOS of 9 days.

```
Out[5]: count    5581.000000
        mean        5.722810
        std         2.506986
        min         1.000000
        25%         4.000000
        50%         5.000000
        75%         9.000000
        max         9.000000
        Name: LOS, dtype: float64
```

Fig 19

From the below bar graph, we can see the Length of stay distribution. A more significant number of patients stayed in the hospital between 0 to 9 days.

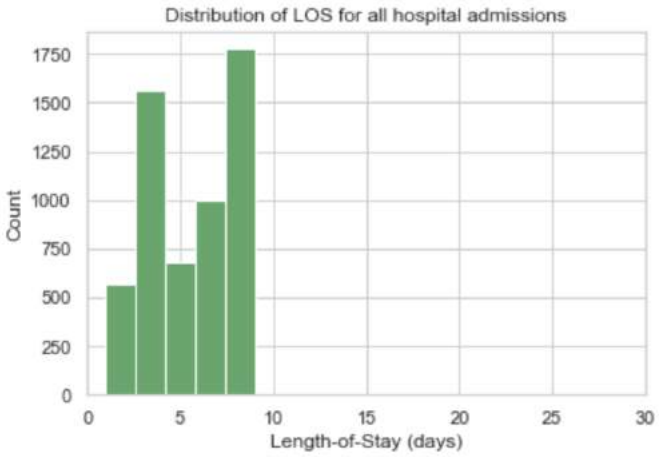

Fig 20

After looking at the LOS distribution, we tried to find which F code contributes more to LOS. The below graph is the comparison of individual F-codes with median Length of stay. It is clear that the F7(Mental Retardation) group of patients stayed longer, followed by F0(Mental Disorders due to psychological conditions)

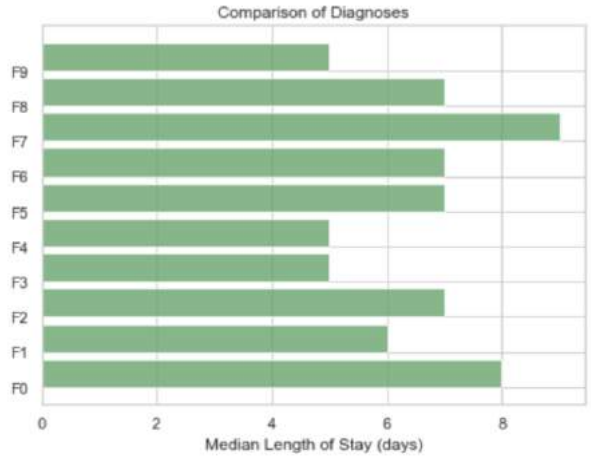

Fig 21

Out of curiosity, we found the top 10 features for predicting LOS. We see that APRDRG_Severity is the ultimate critical feature for predicting LOS, followed by F3(Mood Disorders) and F1(Mental and behavioral disorders due to psychoactive substance).

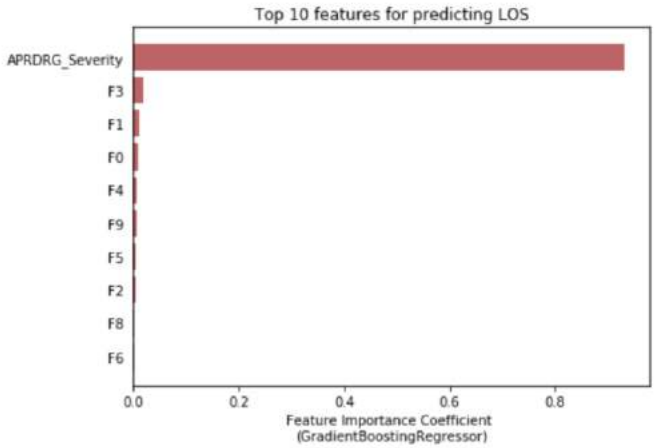

Fig 22

We removed the records which have LOS<=0 as they might affect further analysis. To predict the model, we have split the data into training and test sets in 80:20 proportion, which means now we have 4464 samples in the training set and 1117 samples in the test set. There are many types of regressor models like Gradient boosting regressor, KNearestNeighbour regressor, Random Forest regressor, Linear Regression, SGD Regressor, SVR. We are implementing all these regressor models to predict the model. To decide which models to apply, for each model, R2 and MAE were calculated.

R squared value (R2) is used to define the model's excellent fit, so the more significant the R2 score better the model, and similarly, Mean Absolute Error (MAE) should be less for a good model. We can see that SGD Regressor, Linear Regression, Gradient Boosting Regressor, Random Forest and SVR have good R2 scores from the below figure. SVR has the least MAE score, followed by SGD Regressor, Linear Regression, Gradient Boosting Regressor, Random Forest.

```
R2 scores
SGDRegressor : 0.33631066424430345
GradientBoostingRegressor : 0.3311327487596021
LinearRegression : 0.3362623029884142
KNeighborsRegressor : 0.18300147763045405
RandomForestRegressor : 0.31389682375349615
SVR : 0.3079007927273676

 MAE values
SGDRegressor : 1.7021746603639365
GradientBoostingRegressor : 1.6898821756888585
LinearRegression : 1.6937242187594344
KNeighborsRegressor : 1.8628469113697406
RandomForestRegressor : 1.700999965102717
SVR : 1.6264932537860293
```

Fig 23

R2 Scores

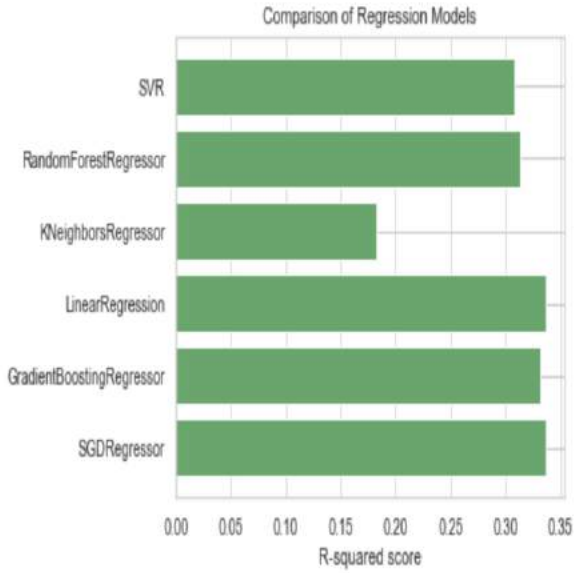

Fig 24

The above figure shows that the SGD regressor has the highest R2 score, followed by Linear Regression and Gradient Boosting Regressor.

MAE Values

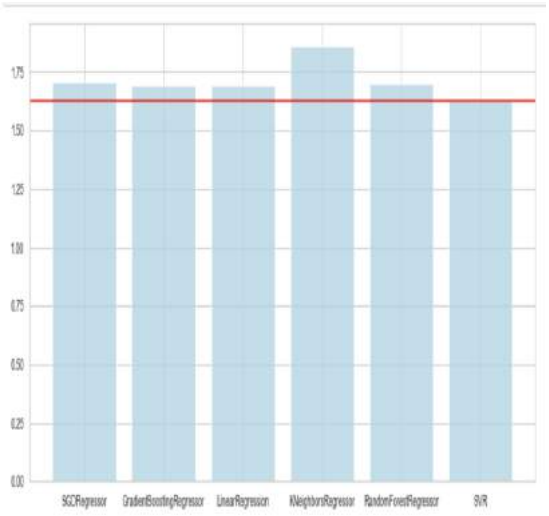

Fig 25

From the above figure, we can see that the SVR model has the least MAE value. So, we implemented all these models to choose the best model to predict LOS.

4.2.1 Gradient Boosting Regression

After predicting important features and obtaining our predicted model, we compared our Actual LOS, Predicted model, median, and Average values. We just applied the expected Gradient Boosting LOS model to the random 20 samples. From the below comparison graph, we see that Actual LOS and Predicted LOS goes hand in hand except at some places. We see that the predicted model LOS at the 9th and 11th records is far greater than the Actual LOS model. This gives a more convoluted picture of the prediction model; in some admissions, it predicts well but not as well in others.

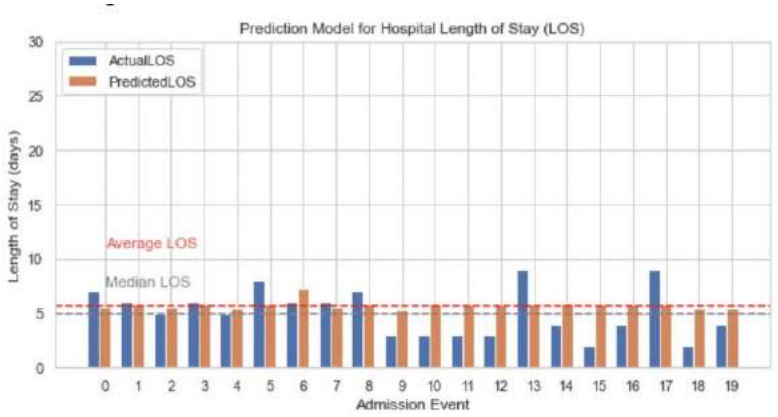

Fig 26

We compare the predicted LOS model with median and average LOS using Root Mean Square error (RMSE) to measure the performance. After running RMSE on all the algorithms, the RMSE for the Prediction model is 0.0730, RMSE for the Median model is 0.07681, and RMSE for the Average model is 0.07374. Our predicted model performs well as we can see that the RMSE for the predicted model is less compared to others. We can also see that the prediction of Length of stay from Gradient Boosting is 2.120 days.

```
Prediction Model days 2.1203138674384494
Median Model days 2.1638316920322294
Average Model days 2.163184593077577
Prediction Model RMS 0.07301756638401131
Median Model RMS 0.07680955265461648
Average Model RMS 0.07373636068724411
```

Fig 27

Below is the graphical representation of RMSE values of the Predicted LOS model (Gradient Boosting), Average LOS model, and Median LOS model. The Gradient Boosting model has lesser LOS, which indicates that the GB regressor is performing well.

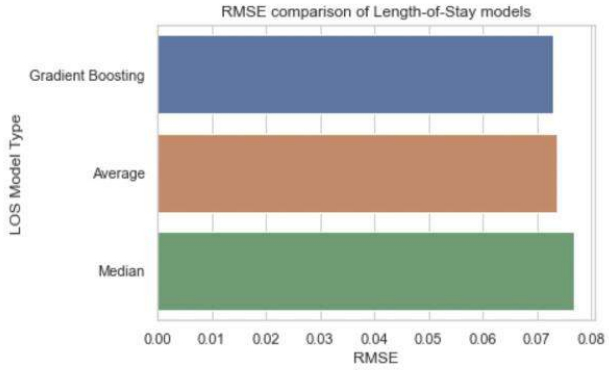

Fig 28

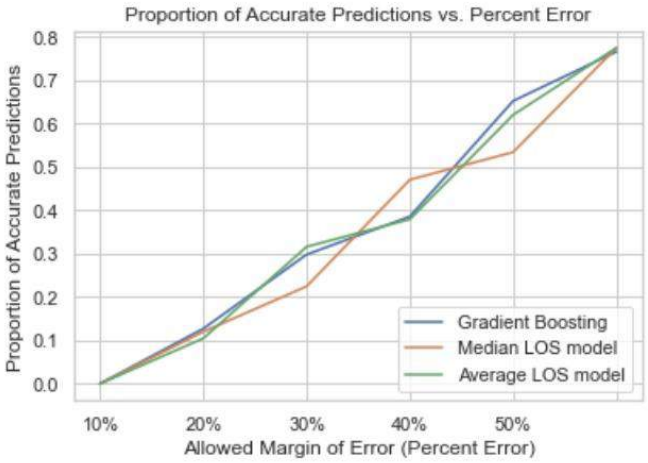

Fig 29

From the above graph, we can conclude that the predicted model (Gradient Boosting) has more accuracy than the Median and Average model with the Allowed margin of error. From the Gradient Boosting model, we can confirm that the predicted model has less RMSE score than other models and less Mean Absolute Error, which is 2.120. The max error score of the model is 5.3241.

4.2.2 Support Vector Regression (SVR)

After predicting the LOS using the SVR model, we applied the model to the random 20 samples. From the below comparison graph, we see that Actual LOS and Predicted LOS goes hand in hand except at some places. We see that at 1st, 9^{th,} and 11th records predicted model LOS is far greater than the Actual LOS model. This gives a more convoluted picture of the prediction model; in some admissions, it portends well but not as well in others.

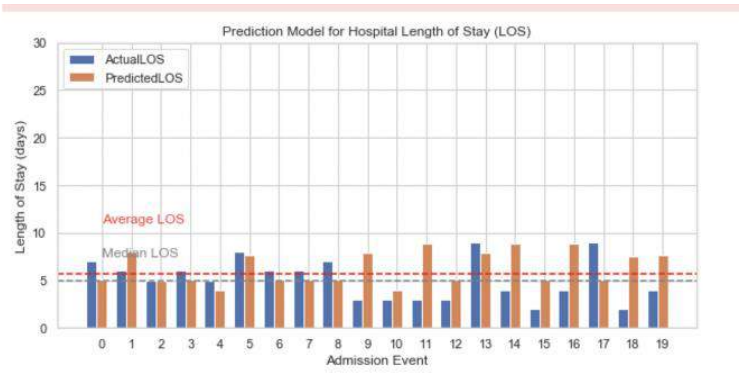

Fig 30

We compare the predicted LOS model with median and average LOS using Root Mean Square error (RMSE) to measure the performance. After running RMSE on all the algorithms, the RMSE for the Prediction model is 0.063399, RMSE for the Median model is 0.07993, and RMSE for the Average model is 0.07624. Our predicted model performs well as we can see that the RMSE for the predicted model is less compared to others. We can also see that the prediction of length of stay from the SVR model is 1.6 days.

```
Prediction Model days 1.626493253786032
Median Model days 2.2820053715308863
Average Model days 2.267122095573982
Prediction Model RMS 0.06339968343441804
Median Model RMS 0.07993881552828291
Average Model RMS 0.07624966411136928
```

Fig 31

Below is the graphical representation of RMSE values of the Predicted LOS model (SVR), Average LOS model, and Median LOS model. We can see that the SVR model has lesser LOS, which is a very good indication that the SVR model performs well.

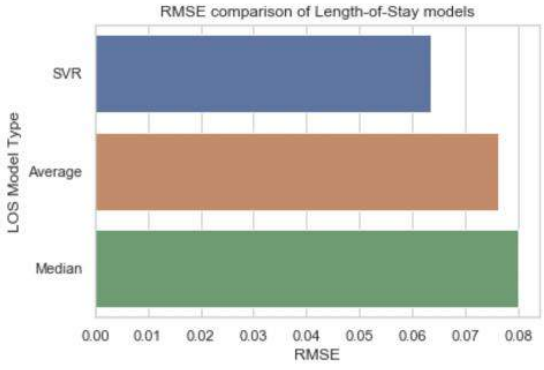

Fig 32

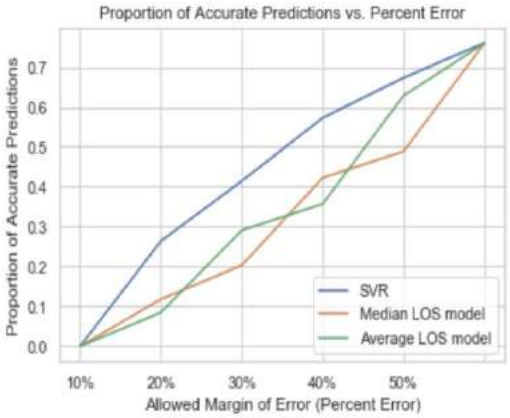

Fig 33

From the above graph, we can conclude that the predicted model (SVR) has more accuracy than the Median and Average model with an Allowed margin of error. From the SVR model, we can confirm that the predicted model has less RMSE score than other models and less Mean Absolute Error, which is 1.626. The max error score of the model is 6.327, and the R2 score is 0.336262.

4.2.3 Linear Regression

After predicting the LOS using the Linear Regression model, we applied the model to the random 20 samples. From the below comparison graph, we see that Actual LOS and Predicted LOS goes hand in hand except at some places. We see that the predicted model LOS at the 9th and 11th records is far greater than the Actual LOS model. This gives a more convoluted picture of the prediction model; in some admissions, it predicts well but not as well in others.

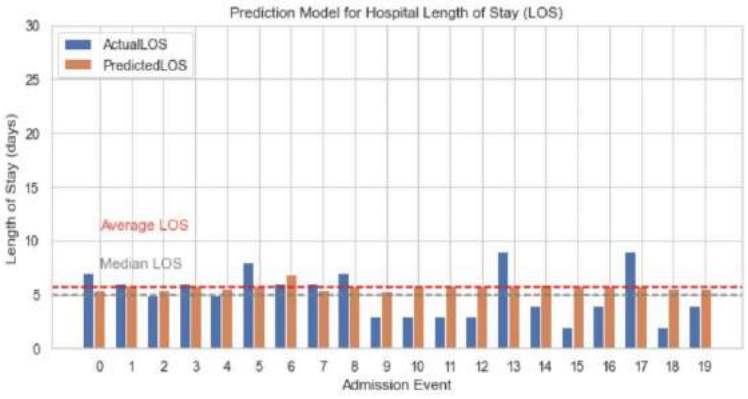

Fig 34

We compare the predicted LOS model with median and average LOS using Root Mean Square error (RMSE) to measure the performance. After running RMSE on all the algorithms, the RMSE for the Prediction model is 0.062087, RMSE for the Median model is 0.079938, and RMSE for the Average model is 0.07624. Our predicted model performs well as we can see that the RMSE

for the predicted model is less compared to others. We can also see that the prediction of length of stay from the SVR model is 1.69 days.

```
Prediction Model days 2.1230593856404303
Median Model days 2.1638316920322294
Average Model days 2.163184593077577
Prediction Model RMS 0.0731221860405339
Median Model RMS 0.07680955265461648
Average Model RMS 0.07373636068724411
```

Fig 35

Below is the graphical representation of RMSE values of the Predicted LOS model (Linear Regression), Average LOS model, and Median LOS model. We can see that the Linear Regression model has lesser LOS, which is an excellent indication that the Linear Regression model performs well.

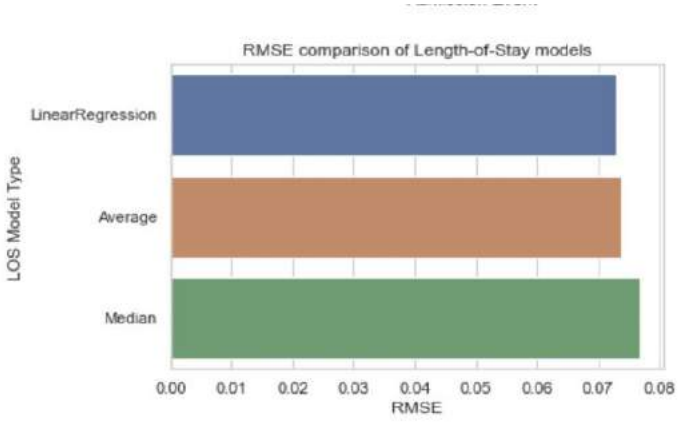

Fig 36

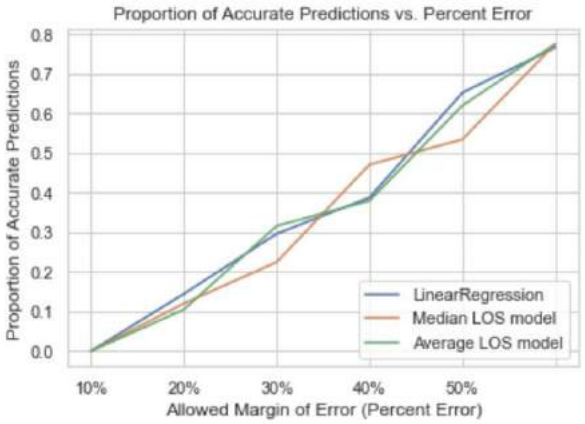

Fig 37

From the above graph, we can conclude that the predicted model (Linear Regression) has more accuracy than the Median and Average model with the Allowed margin of error. From the Linear Regression model, we can confirm that the predicted model has less RMSE score than other models and less Mean Absolute Error, which is 1.6937. The max error score of the model is 5.778, and the R2 score is 0.336262.

4.2.4 Stochastic Gradient Descent Regressor (SGD)

In the below graph, we can see the SGD model and Actual model's comparison applied on random 20 samples. From the below comparison graph, we see that Actual LOS and Predicted LOS goes hand in hand except at some places. We see that the predicted model LOS at the 9th and 11th records is far greater than the Actual LOS model. This gives a more convoluted picture of the prediction model; in some admissions, it predicts well but not as well in others.

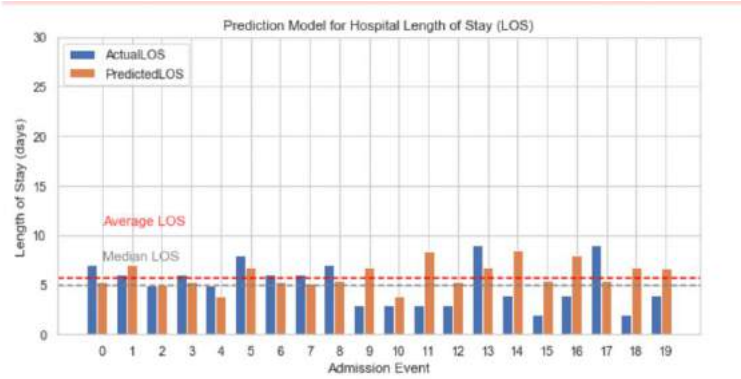

Fig 38

We compare the predicted LOS model with median and average LOS using Root Mean Square error (RMSE) to measure the performance. After running RMSE on all the algorithms, the RMSE for the Prediction model is 0.062242, RMSE for the Median model is 0.079938, and RMSE for the Average model is 0.07624. Our predicted model performs well as we can see that the RMSE for the predicted model is less compared to others. We can also see that the prediction of length of stay from the SGD model is 1.71 days.

```
Prediction Model days 1.7153109062322611
Median Model days 2.2820053715308863
Average Model days 2.267122095573982
Prediction Model RMS 0.06224252488541524
Median Model RMS 0.07993881552828291
Average Model RMS 0.07624966411136928
```

Fig 39

Below is the graphical representation of RMSE values of the Predicted LOS model (SGD), Average LOS model, and Median LOS model. We can see that the SGD model has lesser LOS, which is a very good indication that the SGD model performs well.

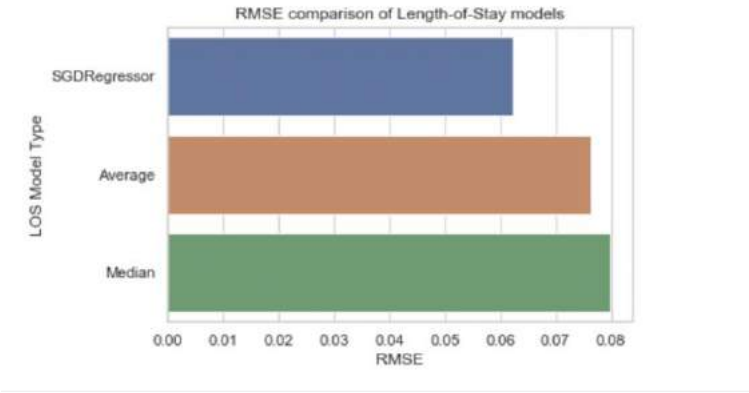

Fig 40

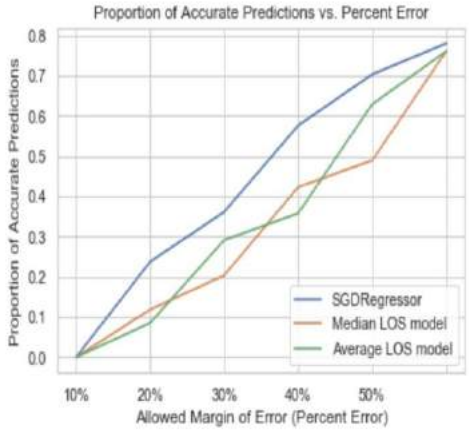

Fig 41

From the above graph, we can conclude that the predicted model (SGD) has more accuracy than the Median and Average model with an Allowed margin of error. From the SGD model, we can confirm that the predicted model has less RMSE score than other models and less Mean Absolute Error, which is 1.6969. The max error score of the model is 5.72, and the R2 score is 0.3377.

4.2.5 Random Forest

In the below graph, we can see the Random Forest model and Actual model's comparison applied on random 20 samples. From the below comparison graph, we see that Actual LOS and Predicted LOS goes hand in hand except at some places. We see that the predicted model LOS at the 9th and 11th records is far greater than the Actual LOS model. This gives a more convoluted picture of the prediction model; in some admissions, it predicts well but not as well in others.

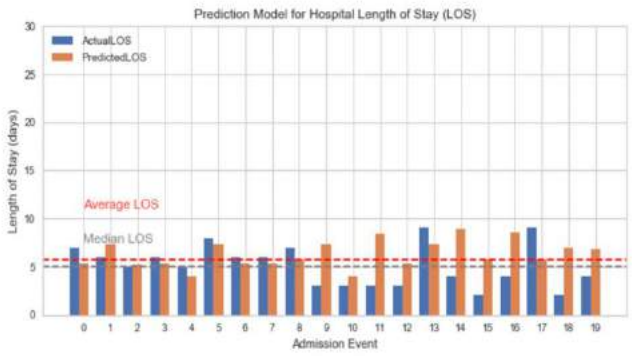

Fig 42

We compare the predicted LOS model with median and average LOS using Root Mean Square error (RMSE) to measure the performance. After running RMSE on all the algorithms, the RMSE for the Prediction model is 0.06312, RMSE for the Median model is 0.079938, and RMSE for the Average model is 0.07624. Our predicted model performs well as we can see that the RMSE for the predicted model is less compared to others. We can also see that the prediction of length of stay from the SGD model is 1.7 days.

```
Prediction Model days 1.7009999651027188
Median Model days 2.2820053715308863
Average Model days 2.267122095573982
Prediction Model RMS 0.06312445308830818
Median Model RMS 0.07993881552828291
Average Model RMS 0.07624966411136928
```

Fig 43

Below is the graphical representation of RMSE values of the Predicted LOS model (Random Forest), Average LOS model, and Median LOS model. We can see that the Random Forest model has lesser LOS, which indicates that the Random Forest model performs well.

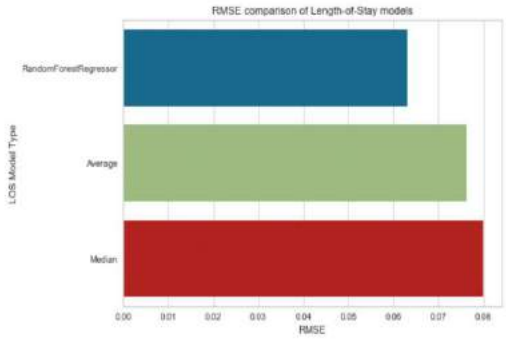

Fig 44

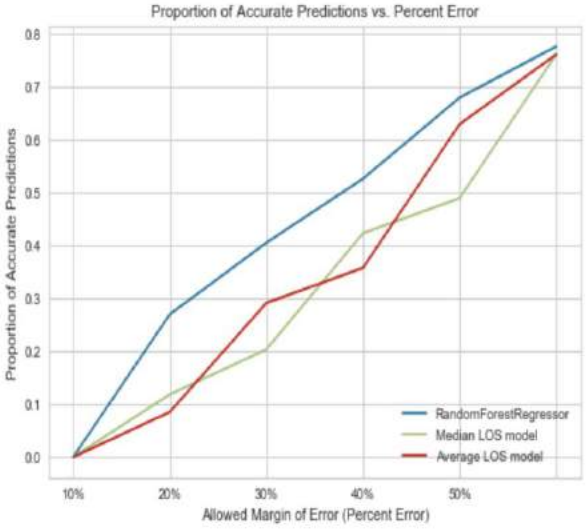

Fig 45

From the above graph, we can conclude that the predicted model (Random Forest) has more accuracy than the Median and Average model with the Allowed margin of error. From the Random Forest model, we can confirm that the predicted model has less RMSE score than other models and less Mean Absolute Error, which is 1.7010. The max error score of the model is 6.0289, and the R2 score is 0.3138.

So, from above all the models, we see that predicted models perform well compared to the median and average models. In all the models' Gradient Boosting model has less Max Error score compared to all models.

We can compare different metrics like R square, RMSE, MAE, MSE, and Max Error on other models we have implemented in the below figure. We see that GBR and SVR have low RMSE, MAE, and MSE and have high R2 values to predict LOS better using these models.

Model	R2	RMSE	MAE	MSE	Max Error
GBR	0.3311	0.0623	1.689	4.339	5.794
SVR	0.3079	0.0633	1.626	4.489	6.327
SGD	0.3377	0.0622	1.696	4.296	5.726
Linear Regression	0.3362	0.0620	1.693	4.305	5.778
Random Forest	0.3138	0.0631	1.701	4.450	6.028

Fig 46

At our sponsors' request, we removed the APRDRG_Severity feature as it is the main feature contributing to LOS and implemented the metrics to see how models are performing. Below we

can see the R square, RMSE, MAE, MSE, Max Error values of different models. The model's performance has gone down after removing the APRDRG_Severity column.

Model	R2	RMSE	MAE	MSE	Max Error
GBR	0.010	0.0758	2.231	6.418	5.259
SVR	-0.015	0.0768	2.218	6.590	5.900
SGD	0.0175	0.0755	2.229	6.373	4.977
Linear Regression	0.015	0.0756	2.229	6.384	4.955
Random Forest	-0.003	0.0763	2.245	6.511	5.6043

Fig 47

As we saw the metrics of different models, excluding APRDRG_Severity, we wanted to see which features contribute to LOS. So, we implemented a few sample variable importance models to know the order of elements. We have performed variable importance plots for Gradient Boosting Regressor and Random Forest Regressor.

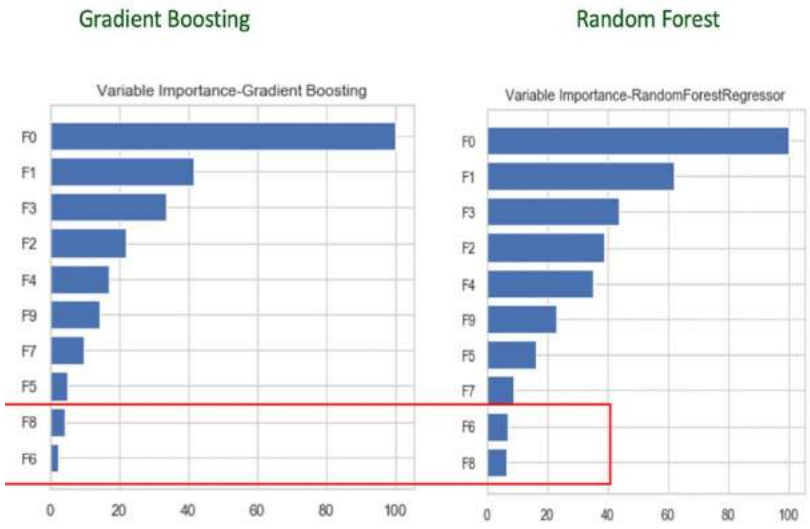

Fig 48

From the above variable importance plots, it is clear that F0 (Mental Disorders due to psychological conditions) is the highest contributor for LOS, followed by F1(Mental and behavioral disorders due to psychoactive substance use) and F3(Mood [affective] disorders). In both the Gradient Boosting model and Random Forest model, the contribution features are in the same order except at F8 and F6. From this, we can conclude that F1 is the highest contribution factor for LOS.

As our Gradient Boosting Regressor and SGD models have given better results out of all the models, we tuned our model in the next step. For tuning the GBR model, we improved the following parameters.

- **N_estimators [300]:** Usually, it defines the number of boosting stages to perform. In other words, it denotes the number of trees in the forest. More number of trees helps in learning the data better. GBR is robust to overfitting, so the number of n-estimators results in better performance. The default value for n_estimators is 100. Here we tuned our model by changing the n_estimators value to 300.
- **Max Depth [200]:** Its function is to limit the number of nodes in the tree. It defines the maximum depth of the decision tree estimator in GBR. To get the model to perform better, we have to find the optimum value of this parameter. We tuned our model by putting the amount as 200.
- **Loss [Huber, Quantile]:** It indicates the loss function to be optimized. It has various functions such as Least square regression (Ls), Least absolute deviation (Lad), Huber, and Quantile. Huber is a combination of Ls and Lad. Quantile allows for quantile regression.
- **Max Features [‘Auto’]:** It defines the number of features to consider while looking for the best split. There are various features in this like auto, sqrt, log2. For auto, max_features=n
  _features. If max_features< n_features, it leads to a reduction in variance and an increase in bias.

  After tuning the GBR model, the RMSE score is 2.436604, and the MAE score is 2.110517. The overall performance of the model has increased by 0.1205.

  For tuning the SGD model, improved the following parameters:

- **Alpha [0.0003]:** The alpha is constant that multiplies the regularization term. Higher the value, the stronger the regularization. It is also used to compute the learning rate when set to ‘Optimal.’ The default value is 0.0001.
- **Loss [‘Huber’]:** Loss function can be used in values like ‘squared_loss,’ ‘Huber,’ ‘epsilon_insensitive.’ The ‘squared_loss’ refers to the ordinary least squares fit. ‘Huber’ modifies squared_loss to focus less on getting outliers correct by switching from squared to linear loss past a distance of epsilon. ‘Epsilon_insensitive’ ignores errors less than epsilon.
- **Learning rate [Adaptive]**: After tuning the above parameters in the SGD model, the RMSE value is 2.455911, and the MAE value is 2.125872. The overall performance of the model has increased by 0.104. Comparing both the tuned GBR and SGD models, GBR has performed very well with low MAE value.

4.2.6 Tensor Flow for Predicting Length of Stay:

The TensorFlow framework provides several added benefits for training a LOS model. First, TensorFlow can run in various environments, whether on CPUs, GPUs, or distributed clusters. This means that the same kind of model can be trained in multiple hospital IT architectures and achieve optimal performance in each. Secondly, with the aid of high-level interfaces, such as

Keras, TensorFlow can model neural network architectures in a very natural way. This enabled us to quickly experiment with different neural network set-ups to find the problem's ideal configuration. Finally, TensorFlow is an actively maintained open-source project, and its performance improves continually through contributions from the open-source machine-learning community.

In the project, we used the Sequential model; as we know in sequential, the layers were connected in sequential order. A sequential model is appropriate for a plain stack of layers where each layer has exactly one layer of input tensor and one layer of the output tensor. Sequential provides training and inference features on the model. The activation used is ReLU (Rectified Linear Unit), as the input is in binary format. The total number of Epochs ran 100 with batch size as 100. The loss function is Mean Absolute Error (MAE), and the metrics are MAE. The optimizer used is Adam. Adam optimization is a stochastic gradient descent method based on the adaptive estimation of first-order and second-order moments. The model is running with the inputs as mentioned earlier.

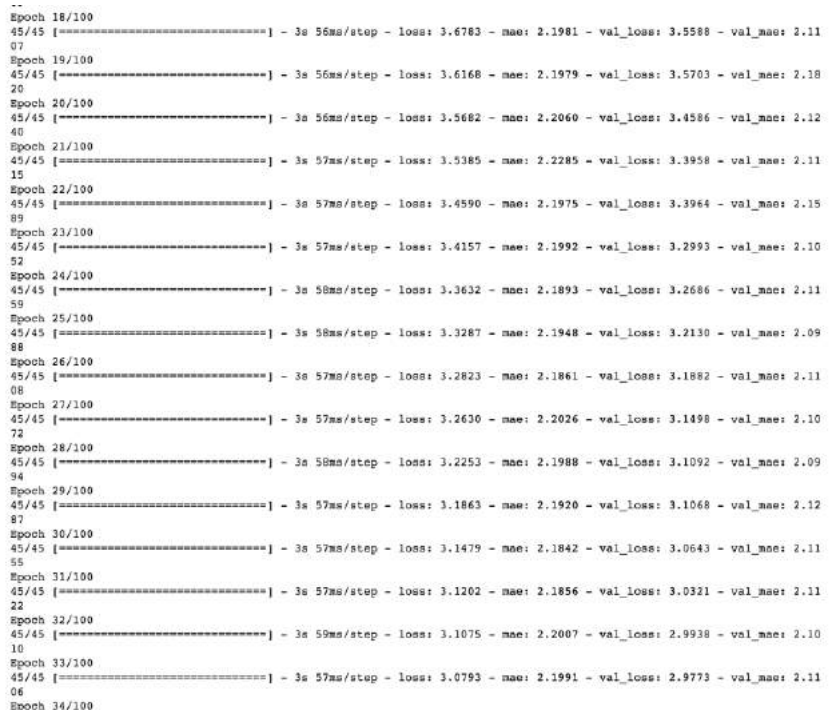

Fig 49

At epoch 25, we got the lowest MAE value, which is 2.09, which is less than the MAE results of GBR and SGD models. Thus, we can say that Tensor Flow has given the best results.

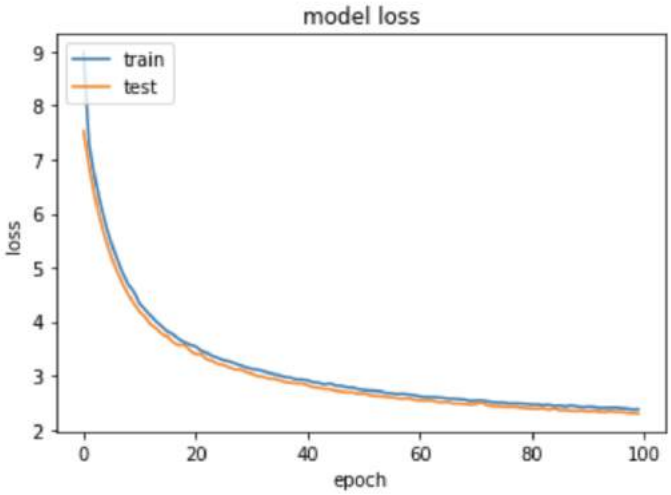

Fig 50

From the above Loss graph plotted against Loss and Epoch for train and test models concerning MAE, we see that both the graphs perfectly align, which means no significant deviation from training and test MAE values excellent.

- **Feature Importance using Permutation Importance:**

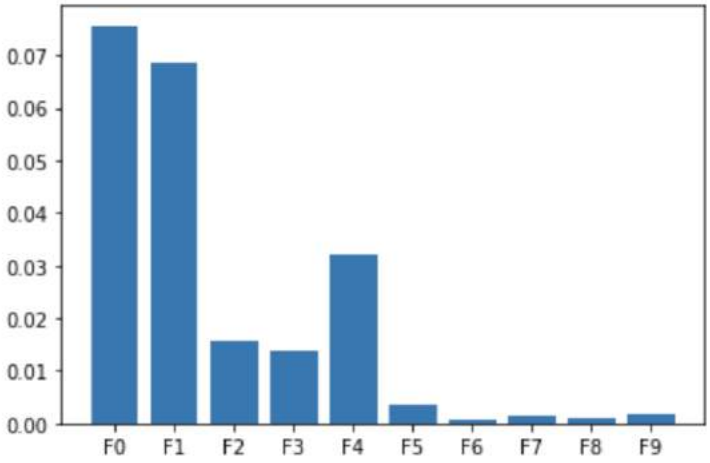

Fig 51

Permutation importance uses models to find feature importance. We calculated the significance after the model is fitted. From the above graph, it is clear that F0 (Mental Disorders due to psychological conditions) and F1(Mental and behavioral disorders due to psychoactive substance use) are the most important features for predicting LOS in Tensor Flow.

Compared with the Gradient Boosting Regressor model and SGD model, Tensor Flow has given better results with low MAE and MSE values. We can conclude that using the tensor flow model can predict the patients' length of stay to predict outcomes better.

4.2.7 Effect of Socio-Economic factors on LOS

From NIS 2016 and 2017 data, prepared a list of socio-economic attributes. Our sponsors are curious to find the effect of these factors on Length of Stay. These categories have been color-coded into four different types. The main aim is to add each group to mental illness codes and find their combined effect on LOS.

- **Socio-Economic Attributes:**

Name	Attribute	Description
Age at admission	AGE	Age in years coded 0-124 years
	AGE_NEONATE	Neonatal age (first 28 days after birth) indicator: (0) non-neonatal age (1) neonatal age
Sex of patient	FEMALE	Indicates gender for NIS beginning in 1998: (0) male, (1) female
Race of patient	RACE	Race, uniform coding: (1) white, (2) black, (3) Hispanic, (4) Asian or Pacific Islander, (5) Native American, (6) other *(For 2016, race contains missing values on about 5 percent of the records.)*
Location of patient's residence	PL_NCHS	Patient Location: NCHS Urban-Rural Code. This is a six-category urban-rural classification scheme for US counties: (1) "Central" counties of metro areas of >=1 million population, (2) "Fringe" counties of metro areas of >=1 million population, (3) Counties in metro areas of 250,000-999,999 population, (4) Counties in metro areas of 50,000-249,999 population, (5) Micropolitan counties, (6) Not metropolitan or micropolitan counties
Median household income for patient's ZIP Code	ZIPINC_QRTL	Median household income quartiles for patient's ZIP Code. For 2016, the median income quartiles are defined as: (1) $1 - $42,999; (2) $43,000 - $53,999; (3) $54,000 - 70,999; and (4) $71,000 or more.
Primary expected payer	PAY1	Expected primary payer, uniform: (1) Medicare, (2) Medicaid, (3) private including HMO, (4) self-pay, (5) no charge, (6) other
Hospital location	HOSP_DIVISION	Census Division of the hospital (STRATA): (1) New England, (2) Middle Atlantic, (3) East North Central, (4) West North Central, (5) South Atlantic, (6) East South Central, (7) West South Central, (8) Mountain, (9) Pacific

Hospital stratifier	NIS_STRATUM	Stratum used to sample hospitals, based on geographic region, control, location/teaching status, and bed size. Stratum information is also contained in the Hospital Weights file.
Hospital characteristics	H_CONTRL	Control/ownership of hospital: (1) government, non-federal, (2) private, non-profit, (3) private, investor-own
	HOSP_LOCTEACH	Location/teaching status of the hospital (STRATA): (1) rural, (2) urban non-teaching, (3) urban teaching
	HOSP_REGION	Region of the hospital: (1) Northeast, (2) Midwest, (3) South, (4) West
	HOSP_DIVISION	Census Division of the hospital (STRATA): (1) New England, (2) Middle Atlantic, (3) East North Central, (4) West North Central, (5) South Atlantic, (6) East South Central, (7) West South Central, (8) Mountain, (9) Pacific
	NIS_STRATUM	Stratum used to sample hospitals beginning in 1998; includes geographic region, control, location/teaching status, and bed size

Table 3

4.2.7.1 Effect of Diagnosis codes, AGE, FEMALE, ZIPINC_QRTL, AGE_NEONATE, RACE on LOS- Model for Socio-Demographic features

Considering the attributes like Age, Female(1 for female, 0 for male), ZIPINC_QRTL(Median household incomes) , AGE_NEONATE(28 days after birth), RACE combined with diagnosis codes, here we are implementing few models like Gradient Boosting regressor, SVR, Linear Regressor, Random Forest.

Model	R2	RMSE	MAE	MSE	Max Error
GBR	0.028	2.510	2.2036	6.302	6.418
SVR	-0.010	2.560	2.243	6.557	5.303
Linear Regression	0.030	2.507	2.203	6.286	5.219
Random Forest	-0.177	2.764	2.322	7.640	7.640

Fig 52

The above figure shows that the GBR and Linear Regression model performed well by giving low MAE values for predicting LOS with diagnosis codes and Socio-demographics features.

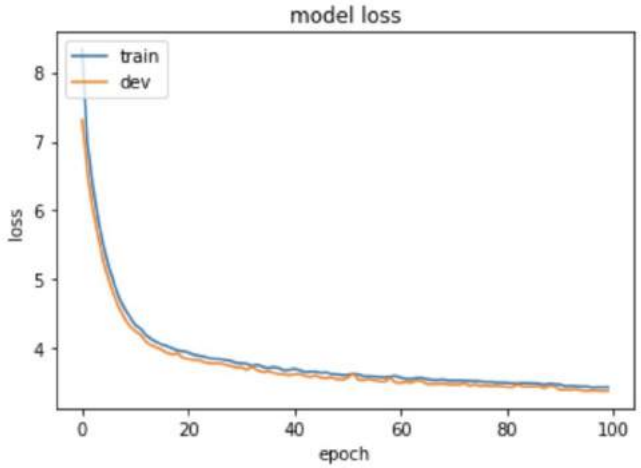

Fig 53

Implementing the loss function graph for the above attributes on train and test set in Tensor Flow on epoch and loss function. Both the graphs align perfectly with each other giving great accuracy. The MAE after implanting tensor flow is 3.25.

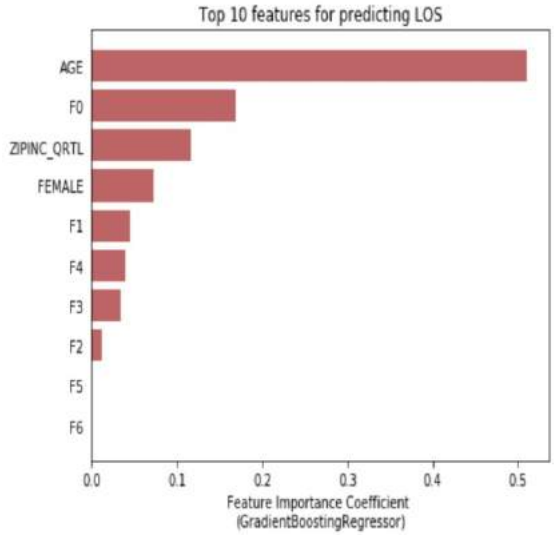

Fig 54

After analyzing the effect of socio-demographic features combined with LOS diagnosis codes, it's time to see which attribute has more impact on LOS. From the above Feature importance graph, it is clear that Age has the highest effect on LOS, followed by F0(Mental disorders due to physiological conditions) and ZIPINC_QRTL (Median household income for patient's ZIP Code). This means that physiological conditions, AGE, and income affect the duration of Length of stay in hospital.

4.2.7.2 Effect of Diagnosis codes, PAY1, HOSP_DIVISION, NIS_STRATUM on LOS- Model for Payer Attributes

Considering the attributes like PAY1, HOSP_DIVISION, NIS_STRATUM combined with diagnosis codes, here we are implementing few models like Gradient Boosting regressor, SVR, Linear Regressor, Random Forest.

Model	R2	RMSE	MAE	MSE	Max Error
GBR	0.042	2.491	2.179	6.209	5.662
SVR	-0.016	2.568	2.242	6.596	5.100
Linear Regression	0.018	2.523	2.220	6.368	5.276
Random Forest	-0.112	2.686	2.269	7.219	7.784

Fig 55

The above figure shows that GBR gave a low MAE value for predicting LOS with diagnosis codes and Payer Information.

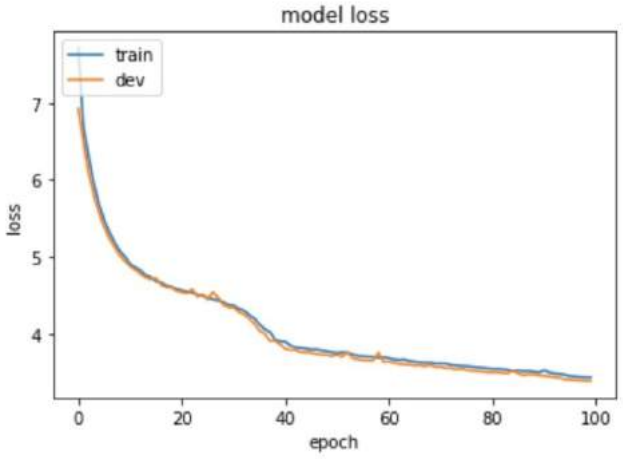

Fig 56

Implementing the loss function graph for the above attributes on train and test set in Tensor Flow on epoch and loss function. Both the graphs align perfectly with each other giving great accuracy. The MAE after implanting tensor flow is 3.191.

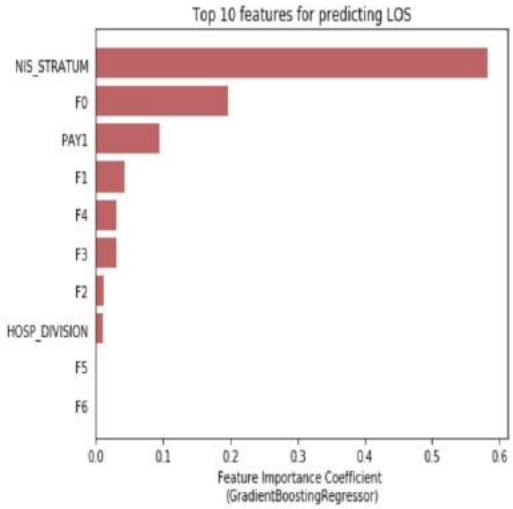

Fig 57

After analyzing the effect of payer features combined with diagnosis codes on LOS, it's time to see which attribute has more impact on LOS. From the above Feature importance graph, it is clear that Stratum, F0(Mental disorders due to physiological conditions) and PAY1(primary expected payer). This means that patients with physiological conditions, type of payer, and stratum affect Length of stay in a hospital.

4.2.7.3 Effect of Diagnosis codes, PL_NCHS, HOSP_DIVISION on LOS- Model for location attributes

Considering the attributes like PL_NCHS, HOSP_DIVISION combined with diagnosis codes, here we are implementing few models like Gradient Boosting regressor, SVR, Linear Regressor, Random Forest.

Model	R2	RMSE	MAE	MSE	Max Error
GBR	0.041	2.493	2.185	6.215	5.596
SVR	0.011	2.532	2.197	6.414	5.208
Linear Regression	0.020	2.520	2.216	6.352	5.227
Random Forest	-0.035	2.592	2.224	6.718	6.768

Fig 58

The above results show that GBR gave a low MAE value for predicting LOS with diagnosis codes and Location attributes.

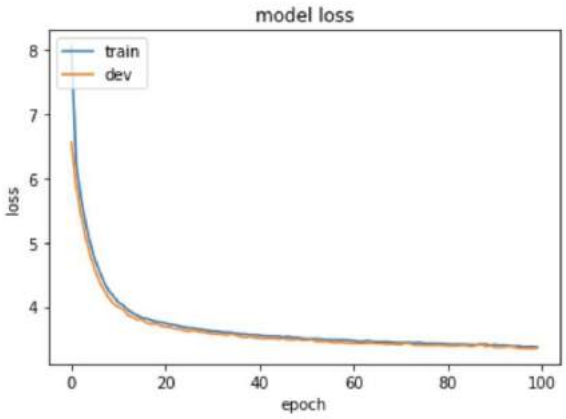

Fig 59

Implementing the loss function graph for the above attributes on train and test set in Tensor Flow on epoch and loss function. Both the graphs align perfectly with each other giving great accuracy. The MAE after implanting tensor flow is 3.25.

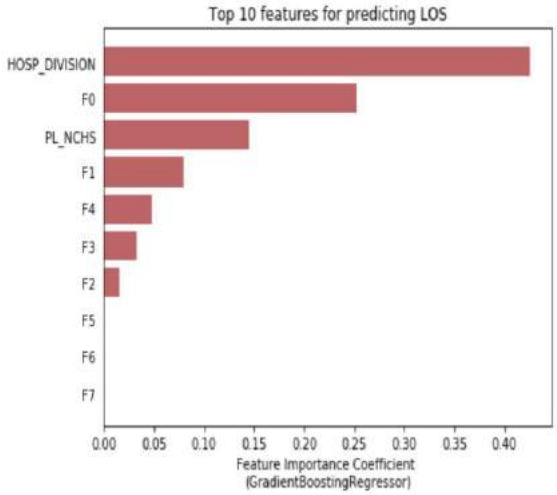

Fig 60

After analyzing the effect of location features combined with diagnosis codes on LOS, it's time to see which attribute has more effect on LOS. From the above Feature importance graph, it is clear that HOSP_DIVISION, F0(Mental disorders due to physiological conditions), and PL_NCHS play a significant role in determining the length of stay.

4.2.7.4 Effect of Diagnosis codes, NIS_STRATUM, H_CONTRL, HOSP_LOCTEACH, HOSP_REGION on LOS- Model for Hospital attributes

Considering the attributes like NIS_STRATUM, H_CONTRL, HOSP_LOCTEACH, HOSP_REGION combined with diagnosis codes, we are implementing a few models like Gradient Boosting regressor, SVR, Linear Regressor, Random Forest.

Model	R2	RMSE	MAE	MSE	Max Error
GBR	0.041	2.493	2.185	6.215	5.596
SVR	0.011	2.532	2.197	6.414	5.208
Linear Regression	0.020	2.520	2.216	6.352	5.227
Random Forest	-0.035	2.592	2.224	6.718	6.768

Fig 61

From the above models, it is clear that GBR gave a low MAE value for predicting LOS with diagnosis codes and Hospital attributes.

Implementing the loss function graph for the above attributes on train and test set in Tensor Flow on epoch and loss function. Both the graphs align perfectly with each other giving great accuracy. The MAE after implanting tensor flow is 3.18.

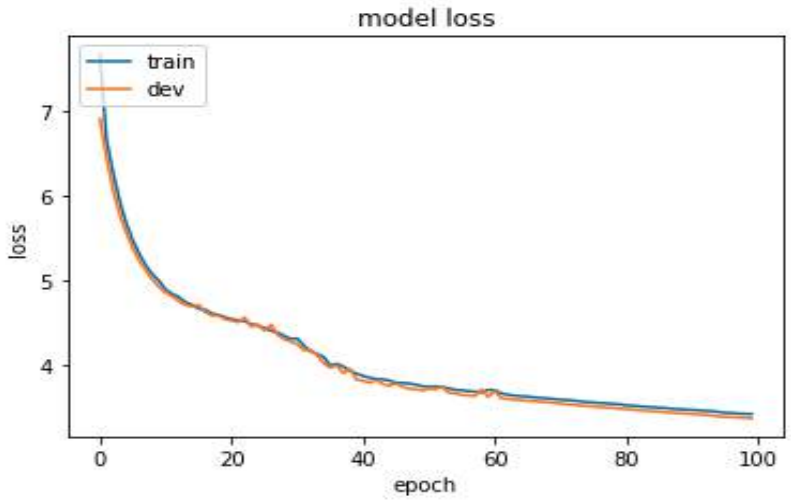

Fig 62

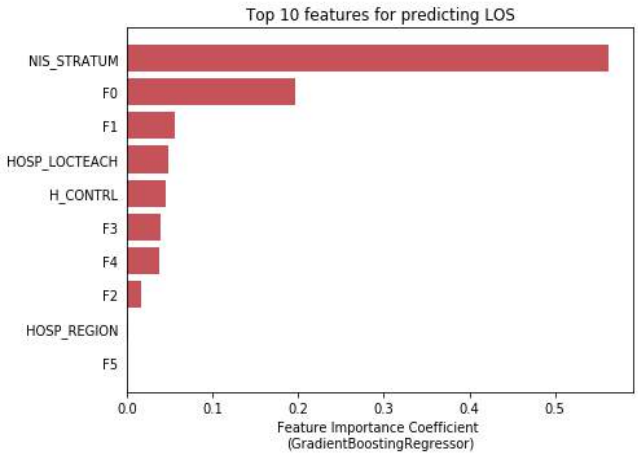

Fig 63

After analyzing the effect of hospital features combined with diagnosis codes on LOS, it's time to see which attribute has more effect on LOS. From the above Feature importance graph, NIS_STATUM, F0(Mental disorders due to physiological conditions) and F1(Mental disorders due to physiological conditions and psychoactive substance use) play an essential role in predicting the length of stay of patients.

4.2.8 Data Distribution Plots

After implementing different algorithms and Tensor flow on Socio-Economic features, it is time to see their data distribution for Length of stay. We tried implementing various factors like Diagnosis codes, HOSP_DIVISION, PAYER, etc., with LOS to see how data is distributed.

4.2.8.1 LOS vs. Diagnosis Codes

Below is the data distribution plot of Length of stay against Diagnosis codes. The X-axis has the count of the number of days the patient spent in the hospital grouped by mean, and Y-axis has the respective diagnosis from F0 to F9.

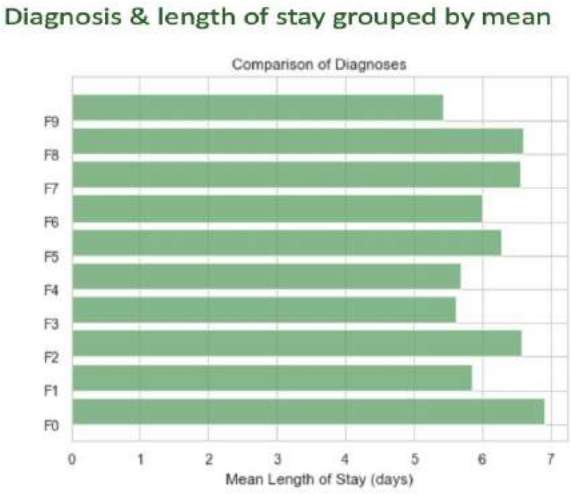

Fig 64

From the above graph, we observe that patients with F0(mental disorders due to physiological conditions) stayed longer in the hospital than other patients, followed by F8 (Pervasive specific developmental disorders)

4.2.8.2 LOS vs. HOSP_DIVISION

Below is the data distribution plot of Length of stay against HOSP_DIVISION. The X-axis has the count of the number of days the patient spent in the hospital grouped by mean. The Y-axis represents Hospital divisions where 1 stands for New England, 2 stands for Middle Atlantic, 3 stands for East North Central, 4 stands for West North Central, 5 stands for South Atlantic, 6 stands for East South Central, 7 for West South Central, 8 for Mountain, 9 for Pacific.

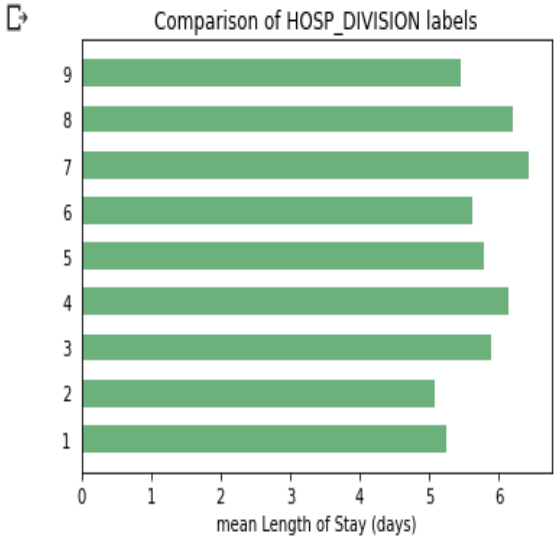

Fig 65

As per the above plot, we see that Hospitals in West south Central-7 and Mountain division-8 had patients who stayed for a longer duration.

4.2.8.3 LOS vs. ZIPINC_QRTL Labels

Below is the data distribution plot of Length of stay against ZIPINC_QRTL. The X-axis has the count of the number of days the patient spent in the hospital group by mean, and Y-axis represents ZIPINC_QRTL, which means house income quartiles for patient's ZIP code where 1 represents mean quartile income in the range of $1- $43,999, 2 stands for $44,000-$55999, 3 stands for $56000-73999 and 4 stands for $74000 or more.

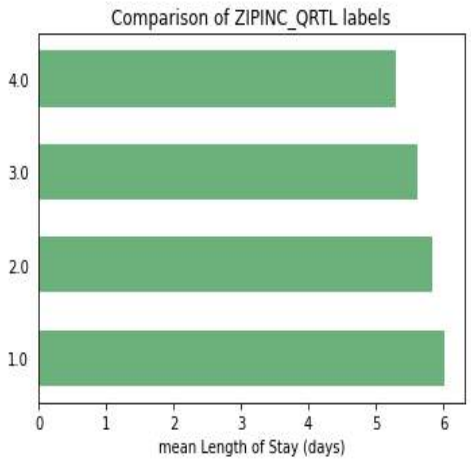

Fig 66

As per the bar graph, we observe that people with median income ($1-$43999) stayed longer in hospital. When we analyzed the patients' primary payer information between $1-$43,999 median income group, we see that most of them belong to payer Medicare.

4.2.8.4 LOS vs. PAY1 Labels

Below is the data distribution plot of Length of stay against PAY1. The X-axis has the count of the number of days the patient spent in the hospital grouped by mean, and Y-axis represents PAY1, which means primary payer where one stands for Medicare, two stands for Medicaid, 3 for Private including HMO, 4 for Self-pay, 5 for no charge, 6 for other.

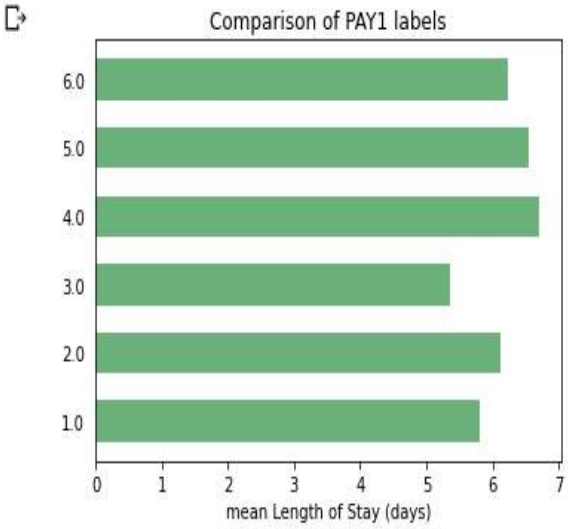

Fig 67

As per the above graph, we see that people who paid self-4 and no charge-5 stayed longer in the hospital.

4.2.8.5 LOS vs PL_NCHS Labels

Below is the data distribution plot of Length of stay against PL_NCHS. The X-axis has the count of the number of days the patient spent in the hospital grouped by mean. Y-axis represents PL_NCHS, which means Patient Location code where one stands for 'Central' counties of metro areas of >=1 million population, two stands for 'Fringe' counties of metro areas of >=1 million population, 3 for Counties in metro areas of 250,000-999,999 people, 4 for Counties in metro areas of 50,000-249,999 people, 5 for Micropolitan counties, 6 for Not metropolitan or micropolitan counties.

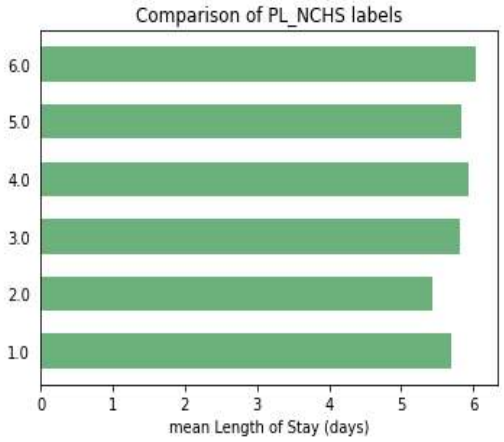

Fig 68

As per the bar graph, we observe there is no much difference between different locations. The Average Length of stay is more than 5.5 days for all locations.

4.2.8.6 LOS vs. HOSP_LOCTEACH

Below is the data distribution plot of Length of stay against HOSP_LOCTEACH. The X-axis has the count of the number of days the patient spent in the hospital grouped by mean, and Y-axis represents HOSP_LOCTEACH, which means Loc/ Teaching status of the hospital where one stands for Rural, 2 for Urban non-teaching, 3 for Urban teaching.

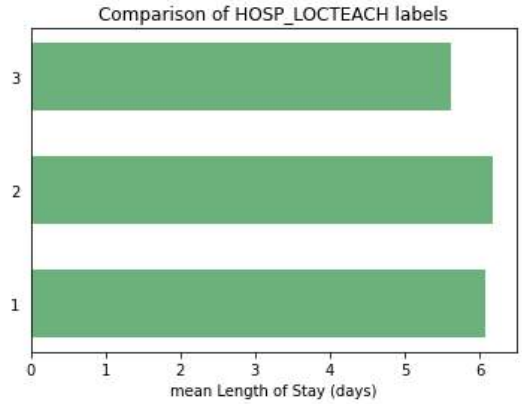

Fig 69

As per the bar graph, we observe that patients in Urban non-teaching and rural stayed Longer.

4.2.8.7 LOS vs. H_CONTROL Labels

Below is the data distribution plot of Length of stay against H_CONTROL. The X-axis has the count of the number of days the patient spent in the hospital grouped by mean, and Y-axis represents Control/Ownership of the Hospital where one stands for Government, non-federal, 2 for Private, non-profit, 3 for Private, investor own.

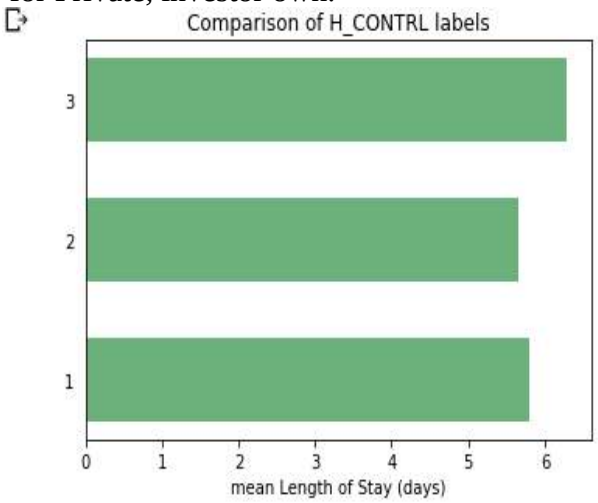

Fig 70

As per the bar graph, we observe that patients stayed longer in Private, investor-own hospitals than non-profit and Government hospitals.

4.2.8.8 LOS vs. HOSP_REGION Labels

Below is the data distribution plot of Length of stay against HOSP_REGION. The X-axis has the count of the number of days the patient spent in the hospital grouped by mean, and Y-axis represents the hospital's region where 1 stands for Northeast, 2 for Midwest, 3 for South, 4 for West.

As per the bar graph, we observe that patients stayed for less duration in hospital in the Northeast Region and more in the Midwest and South.

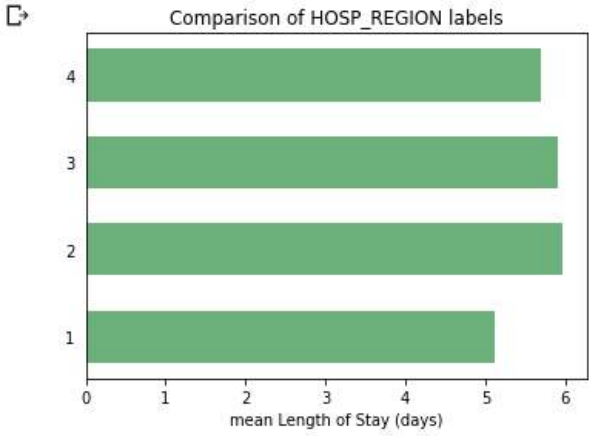

Fig 71

4.2.8.9 LOS vs. AGE

Below is the data distribution plot of Length of stay against AGE. The X-axis has the count of the number of days the patient spent in the hospital grouped by mean, and Y-axis represents the Age of the patient.

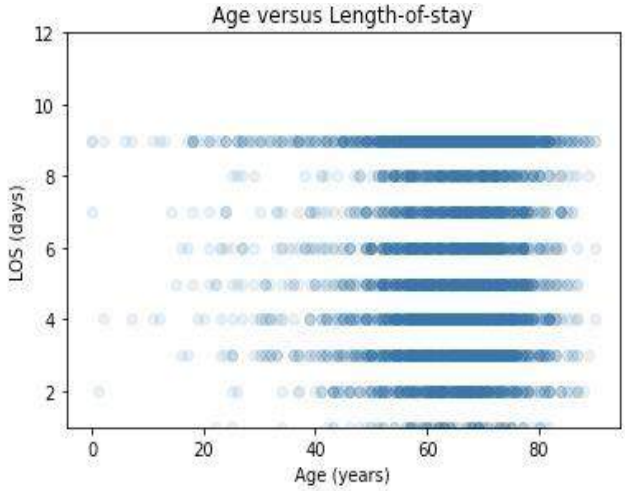

Fig 72

As per the graph, we observe that patients between 45-80 stayed longer in hospital between 8-10 days compared with other age groups.

4.2.9 K Fold Cross-Validation

Cross-validation is a resampling technique used to evaluate machine learning models on a limited data sample. The procedure has a parameter called k, which refers to the number of groups that a given sample needs to be split. Initially, the entire data is randomly divided into k folds and then fit the model using k-1 folds and validate on the kth fold. Repeat this process until every k fold serves as the test set. Then take the average of recorded performances. That will be the performance metric for the model.

We implemented k fold validation to check the model's accuracy. As the Gradient boosting regressor has given the best results for all features compared to other models, the model performance was evaluated using cross-validation.

4.2.9.1 Cross-Validation Results for features F0-F9

Cross-validation is implemented on the GBR model for features F0-F9. The k-fold split chosen is 4. As the name implies, negative MAE is simply the negative of the MAE. Since MAE is an error metric, i.e., the lower, the better, whereas the negative MAE is the opposite, in the below table, we can compare RMSE and MAE values for the GBR and K fold cross-validation model. The Neg_RMSE is -2.496, and Neg_MAE is -2.176. Both the metrics for different implementations are the same, which says that the model has performed well.

Model	RMSE	MAE
80:20	2.436	2.110
Model	**Neg_RMSE**	**Neg_MAE**
K fold Cross validation	-2.496	-2.176

Fig 73

4.2.9.2 Cross-Validation Results for features F0-F9 and AGE, FEMALE, ZIPINC_QRTL, AGE_NEONATE, RACE

Cross-validation is implemented on the GBR model for F0-F9 and AGE, FEMALE, ZIPINC_QRTL, AGE_NEONATE, and RACE. The k-fold split chosen is 4. We can compare RMSE and MAE values for the GBR model and K fold cross validation model in the below table. The Neg_RMSE is -2.480, and Neg_MAE is -2.160. Both the metrics for different implementations are the same, which says that the model has performed well.

Model	RMSE	MAE
80:20	2.506	2.127
Model	**Neg_RMSE**	**Neg_MAE**
K fold Cross validation	-2.480	-2.160

Fig 74

4.2.9.3 Cross-Validation Results for features F0-F9 and PAY1, HOSP_DIVISION, NIS_STRATUM

Cross-validation is implemented on the GBR model for features F0-F9 and PAY1, HOSP_DIVISION, NIS_STRATUM. The k-fold split chosen is 4. We can compare RMSE and MAE values for the GBR model and K fold cross-validation model in the below table. The Neg_RMSE is -2.488, and Neg_MAE is -2.128. Both the metrics for different implementations are the same, which says that the model has performed well.

Model	RMSE	MAE
80:20	2.410	2.071
Model	**Neg_RMSE**	**Neg_MAE**
K fold Cross validation	-2.488	-2.128

Fig 75

4.2.9.4 Cross-Validation Results for features F0-F9 and PL_NCHS, HOSP_DIVISION

Cross-validation is implemented on the GBR model for features F0-F9 and PL_NCHS, HOSP_DIVISION. The k-fold split chosen is 4. We can compare RMSE and MAE values for the GBR model and K fold cross-validation model in the below table. The Neg_RMSE is -2.473, and Neg_MAE is -2.146. Both the metrics for different implementations are the same, which says that the model has performed well.

Model	RMSE	MAE
80:20	2.410	2.071
Model	**Neg_RMSE**	**Neg_MAE**
K fold Cross validation	-2.473	-2.146

Fig 76

4.2.9.5 Cross-Validation Results for features F0-F9 and NIS_STRATUM, H_CONTRL, HOSP_LOCTEACH, HOSP_REGION

Cross-validation is implemented on the GBR model for features F0-F9 and NIS_STRATUM, H_CONTRL, HOSP_LOCTEACH, HOSP_REGION. The k-fold split chosen is 4. We can compare RMSE and MAE values for the GBR model and K fold cross-validation model in the below table. The Neg_RMSE is -2.458, and Neg_MAE is -2.131. Both the metrics for different implementations are the same, which says that the model has performed well.

Model	RMSE	MAE
80:20	2.424	2.064
Model	**Neg_RMSE**	**Neg_MAE**
K fold Cross validation	-2.458	-2.131

Fig 77

4.2.10 Predicting LOS – Final Model

4.2.10.1 Model Performance for features F0-F9

The below table shows the comparisons of different models like GBR, SVR, and SGD. Linear Regression and Random Forest on various metrics like R2, RMSE, MAE, MSE, Max error for features F0-F9. We see that both GBR and Linear regression performed well by giving low errors.

Model	R2	RMSE	MAE	MSE	Max Error
GBR	0.010	2.533	2.231	6.418	5.259
SVR	-0.015	2.567	2.218	6.590	5.900
SGD	0.0175	2.524	2.229	6.373	4.977
Linear Regression	0.015	2.526	2.229	6.384	4.955
Random Forest	-0.003	2.551	2.245	6.511	5.6043

Fig 78

The features mentioned above are implemented even in Tensor Flow. In the below graph, we can see the loss graph for train and dev sets. Both the graphs go inline, which is a good indicator of accuracy. The MAE is 2.89.

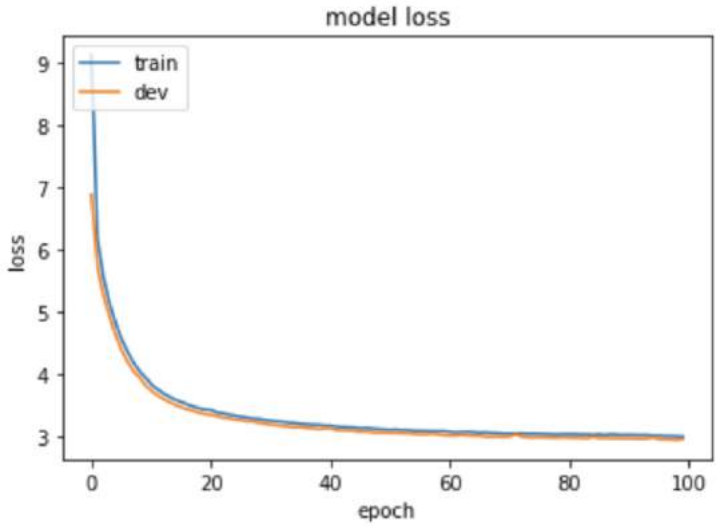

Fig 79

In the below feature importance plot for predicting LOS depending on features from F0-F9, it is clear that F0(Mental disorders due to physiological conditions) is the top feature contributing to LOS, followed by F4 and F1. It is understandable that patients with mental disorders were affected much and stayed for a more extended hospital period.

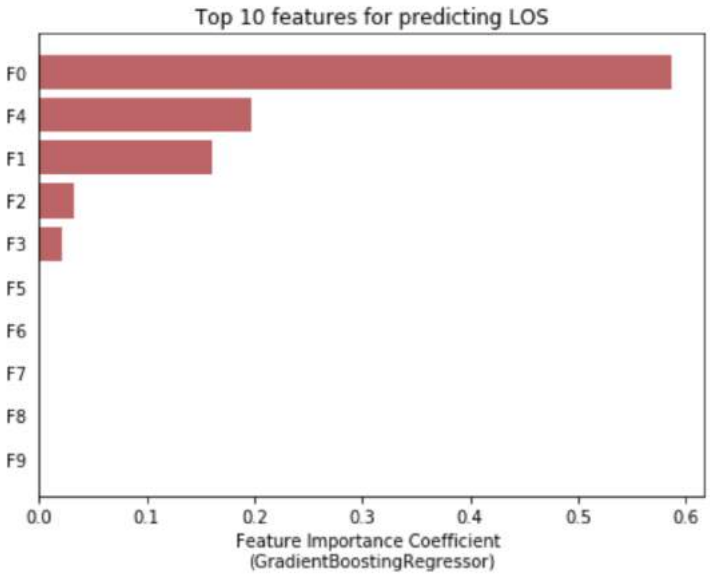

Fig 80

4.2.10.2 Model Performance for features F0-F9 and AGE, FEMALE, ZIPINC_QRTL, LOS, PAY1, HOSP_LOCTEACH

The below table shows the comparisons of different models like GBR, SVR, and SGD. Linear Regression and Random Forest on various metrics like R2, RMSE, MAE, MSE, Max error for features F0-F9. We see that both GBR and Linear regression performed well by giving low errors.

Model	R2	RMSE	MAE	MSE	Max Error
GBR	0.047	2.485	2.172	6.177	6.489
SVR	-0.009	2.558	2.240	6.548	5.329
Linear Regression	0.0353	2.501	2.190	6.258	5.468
Random Forest	-0.154	2.736	2.283	7.489	7.328

Fig 81

The features mentioned above are implemented even in Tensor Flow. In the below graph, we can see the loss graph for train and dev sets. Both the graphs go inline, which is a good indicator of accuracy. The MAE is 3.26.

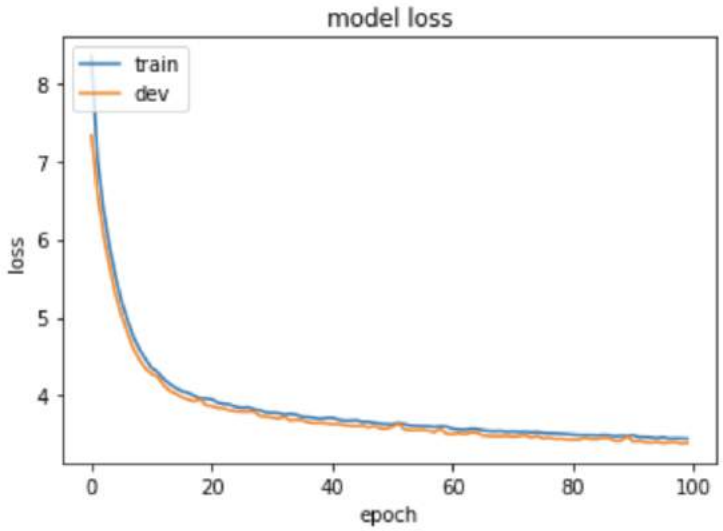

Fig 82

In the below feature importance plot for predicting LOS, it is clear that AGE is the most important factor contributing to LOS, followed by ZIPINC_QRTL and PAY1. Understandably, Aged people will stay longer in hospitals.

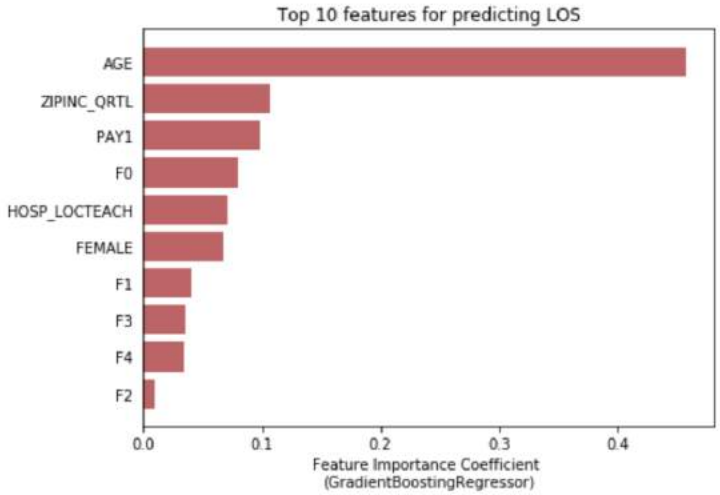

Fig 83

4.3 Problem Statement 3- Predicting Total Charges

The primary aim is to predict Total charges for patients admitted to the hospitals because of lobectomy and mental illness. In this module, the findings will be what attributes affect the total charges and the amount of money paid by the patient while leaving the hospital.

4.3.1. Distribution of TOTCHG

TOTCHG is the attribute for Total Charges, which is directly proportional to the length of stay.

```
count      5581.000000
mean      80336.271636
std       25357.353719
min        5977.000000
25%       59092.000000
50%       86195.000000
75%      105111.000000
max      105111.000000
Name: TOTCHG, dtype: float64
```

Fig 84

4.3.2 Data Distribution Plots

4.3.2.1 Comparison of Diagnosis codes

Below is the data distribution plot of Total Charges against Diagnostic codes. The X-axis is scaled to mean total charges, whereas Y-axis consists of diagnostic codes from F0 to F9. People affected by F0(Mental Disorders due to physiological conditions), F2(Schizophrenia, Schizotypal, delusional, and other non-mood psychotic disorders) paid more charges compared to other diagnostic codes followed by F8 and F6.

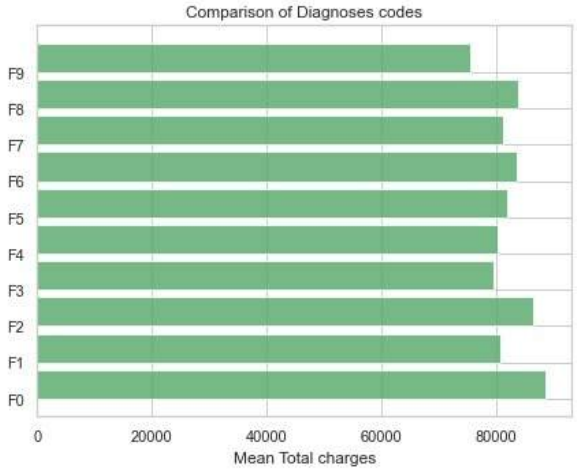

Fig 85

4.3.2.2 Comparison of HOSP_REGION labels

Below is the data distribution plot of Total Charges against Hospital Regions. The X-axis is scaled to mean total charges. Y-axis consists of Hospital regions where 1 represents Northeast, 2 represents Midwest, 3 represents South, and 4 represents West region. Patients in West region-4 paid more charges than compared to other regions.

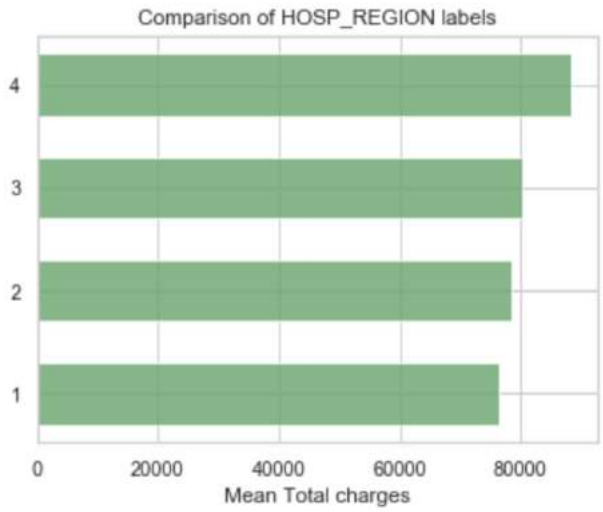

Fig 86

4.3.2.3 Comparison of HOSP_DIVISION labels

Below is the data distribution plot of Total Charges against Hospital Division labels. The X-axis is scaled to mean total charges. Y-axis represents Hospital divisions where 1 stands for New England, 2 stands for Middle Atlantic, 3 stands for East North Central, 4 stands for West North Central, 5 stands for South Atlantic, 6 stands for East South Central, 7 for West South Central, 8 for Mountain, 9 for Pacific. Patients in Mountain and Pacific region paid more charges than compared people from the other areas.

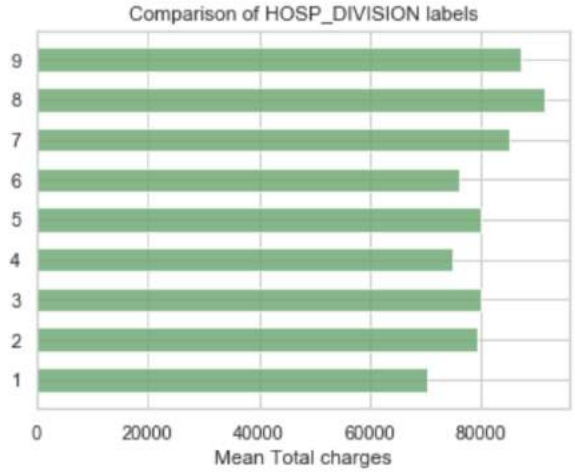

Fig 87

4.3.2.4 Comparison of ZIPINC_QRTL labels

Below is the data distribution plot of Total Charges against ZIPINC_QRTL labels. The X-axis is scaled to mean total charges. Y-axis represents ZIPINC_QRTL, which implies house income quartiles for patients ZIP code where 1 represents mean quartile income in the range of $1-$43,999, 2 stands for $44,000-$55999, 3 stands for $56000-73999 and 4 stands for $74000 or more.

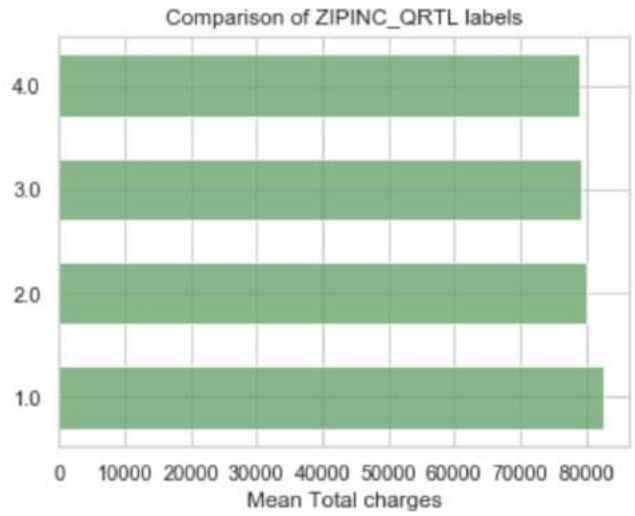

Fig 88

The above graph shows that patients with an income range between $1-$43k paid more than other groups. The primary payer for that median income group is Medicare.

4.3.2.5 Comparison of PL_NCHS labels

Below is the data distribution plot of Total Charges against PL_NCHS labels. The X-axis is scaled to mean total charges. Y-axis represents PL_NCHS. The X-axis has the count of the number of days the patient spent in the hospital grouped by mean. Y-axis represents PL_NCHS, which means Patient Location code where 1 stand for 'Central' counties of metro areas of >=1 million population, 2 stands for 'Fringe' counties of metro areas of >=1 million population, 3 for Counties in metro areas of 250,000-999,999 population, 4 for Counties in metro areas of 50,000-249,999 population, 5 for Micropolitan counties, 6 for Not metropolitan or micropolitan counties.

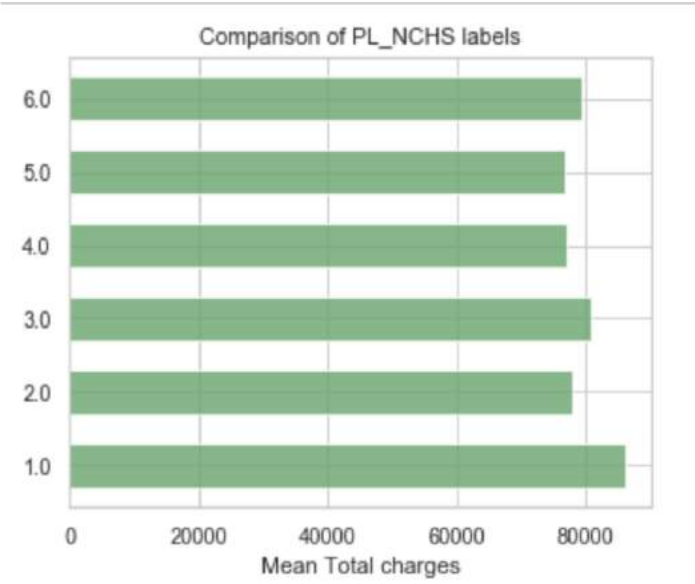

Fig 89

From the above graph, it is clear that patients in Central countries of Metro areas >=1 million population paid more charges, followed by Non-metropolitan counties.

4.3.2.6 Comparison of PAY1 labels

Below is the data distribution plot of Total Charges against PAY1. The X-axis has the Total charges grouped by mean, and Y-axis represents PAY1, which means primary payer where 1 stands for Medicare, 2 stands for Medicaid, 3 for Private including HMO, 4 for Self-pay, 5 for no charge, 6 for other. From the below graph, it is clear that category-No charge paid more charges followed by Self-pay category.

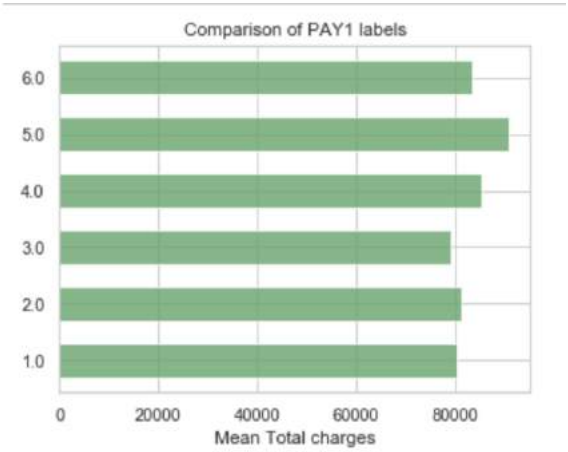

Fig 90

4.3.2.7 Comparison of HOSP_LOCTEACH labels

Below is the data distribution plot of Total Charges against HOSP_LOCTEACH. The X-axis has the Total charges grouped by mean, and Y-axis represents HOSP_LOCTEACH, which means Loc/ Teaching status of the hospital where 1 stand for Rural, 2 for Urban non-teaching, 3 for Urban teaching. The below graph shows that patients in the Urban Non-teaching location paid more than other areas.

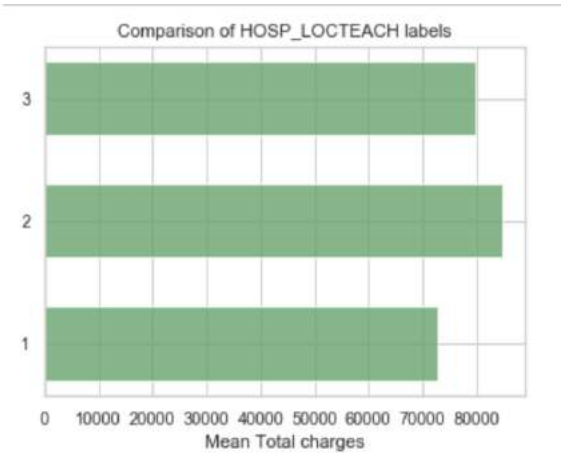

Fig 91

4.3.2.8 Comparison of Age labels

Below is the data distribution plot of Total Charges against PAY1. The X-axis has the Total charges grouped by mean, and Y-axis represents the Age of the patients. The graph shows that patients between the age group 60-80 paid more expenses as they stayed longer in the hospital.

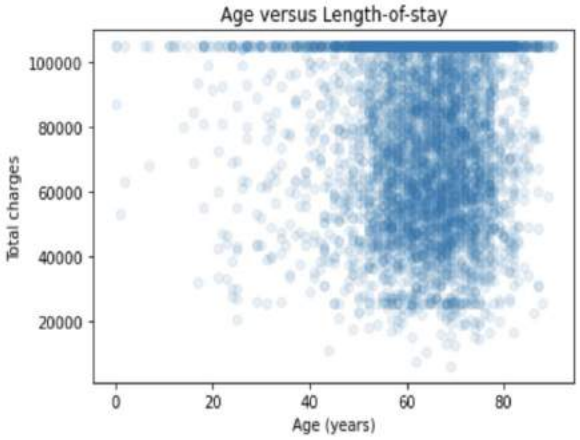

Fig 92

4.3.3 Predicting Total Charges

4.3.3.1 Model performance for features F0-F9

Implementation of different models to predict Total charges include implementing other models like Gradient Boosting Regressor, Support Vector Regressor, Linear Regression, Random Forest. Various metrics like R square, RMSE, MAE, MSE, Max Error, are being estimated. From the below table, it is clear that Linear regression has performed well by given low error values followed by GBR. Both GBR and Linear Regression have shown a short MAE of 0.323, and Linear regression has provided a low RMSE of 0.390. We can conclude that linear regression has performed better.

Model	R2	RMSE	MAE	MSE	Max Error
GBR	-0.000	0.392	0.323	0.153	2.535
SVR	-0.095	0.4105	0.308	0.168	2.667
Linear Regression	0.006	0.390	0.323	0.152	2.532
Random Forest	-0.007	0.393	0.324	0.155	2.536

Fig 93

The features mentioned above are implemented even in Tensor Flow. In the below graph, we can see the loss graph for train and dev sets. Both the graphs go inline, which is a good indicator of accuracy. The MAE is 3.15. But compared to Tensor flow, Linear regression and GBR has given low MAE.

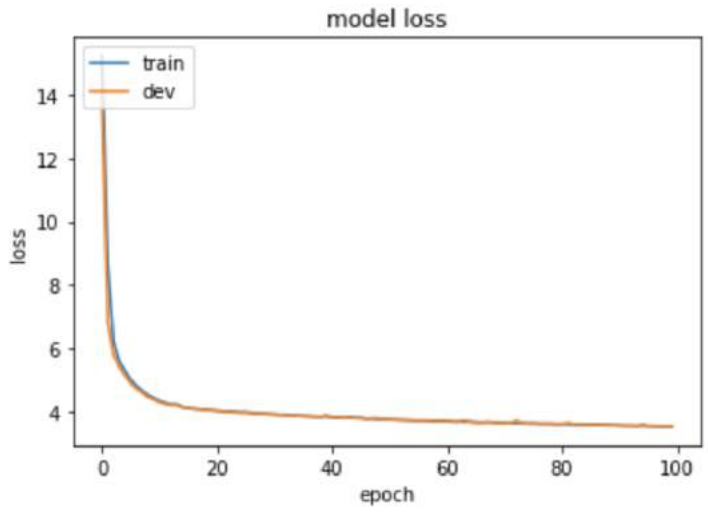

Fig 94

Cross-Validation results for features F0-F9

Cross-validation is implemented on the GBR model for features F0-F9. The k-fold split chosen is 4. As the name implies, negative MAE is simply the negative of the MAE. Since MAE is an error metric, i.e., the lower, the better, whereas the negative MAE is the opposite, in the below table, we can compare RMSE and MAE values for the GBR model and K fold cross-validation model. The Neg_RMSE is -0.392, and Neg_MAE is -0.320. Both the metrics for different implementations are the same, which says that the model has performed well.

Model	RMSE	MAE
80:20	0.381	0.310
Model	**Neg_RMSE**	**Neg_MAE**
K fold Cross validation	-0.392	-0.320

Fig 95

Feature Importance Plot

In the below feature importance plot for predicting Total Charges depending on features from F0-F9, it is clear that F0(Mental disorders due to physiological conditions) is the top feature contributing to Total Charges, followed by F4 and F2. It is understandable that patients with mental disorders are affected much and stayed for a more extended period in the hospital, so they have to pay more charges.

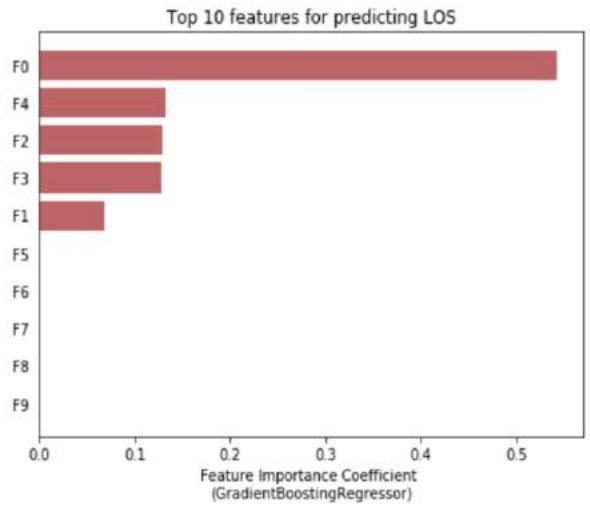

Fig 96

4.3.3.2 Model performance for Socio-Demographic features

Different models like GBR, SVR, Linear Regression, Random Forest are implemented on various Socio-Demographic features to predict Total Charges. The features include Age, Sex, Race, Median household income for patient's ZIP code, primary payer, HOSP_LOCTEACH and HOSP_REGION. The below table shows that Linear regression has performed well by given low error values followed by GBR. Both GBR and Linear Regression have a low MAE of 0.261, and Linear regression and GBR have shown a low RMSE of 0.346. We can conclude that linear regression has performed better by giving a low Max Error.

Model	R2	RMSE	MAE	MSE	Max Error
GBR	0.217	0.346	0.261	0.120	2.776
SVR	0.202	0.350	0.255	0.122	2.806
Linear Regression	0.218	0.346	0.261	0.120	2.791
Random Forest	0.106	0.370	0.278	0.137	2.815

Fig 97

The features mentioned above are implemented even in Tensor Flow. In the below graph, we can see the loss graph for train and dev sets. Both the graphs go inline, which is a good indicator of accuracy. The MAE is 4.08. But compared to Tensor flow, Linear regression and GBR has given low MAE.

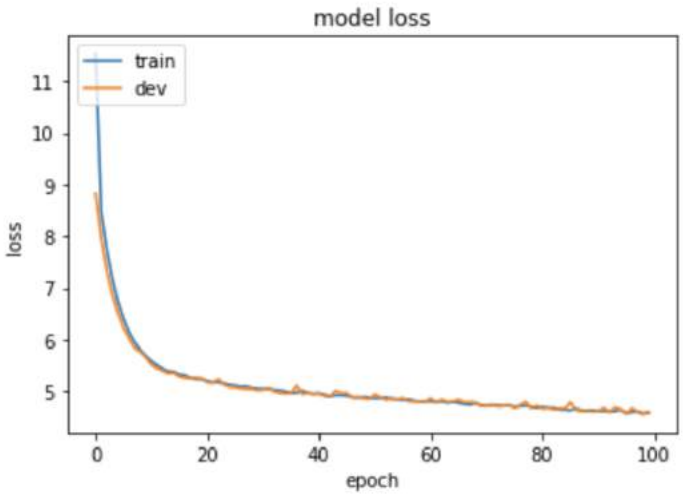

Fig 98

Cross-Validation results for Socio-Demographic features

Cross-validation is implemented on the GBR model for Socio-Demographic features. The k-fold split chosen is 4. As the name implies, negative MAE is simply the negative of the MAE. Since MAE is an error metric, i.e., the lower, the better, whereas the negative MAE is the opposite, in the below table, we can compare RMSE and MAE values for the GBR model and K fold cross-validation model. The Neg_RMSE is -0.347, and Neg_MAE is -0.261. Both the metrics for different implementations are the same, which says that the model has performed well.

Model	RMSE	MAE
80:20	0.343	0.258
Model	**Neg_RMSE**	**Neg_MAE**
K fold Cross validation	-0.347	-0.261

Fig 99

Feature Importance Plot

From the below feature importance plot implemented on Socio-Demographic features for predicting Total Charges, it is clear that AGE is the top feature. even this is easily predictable because older people stay for a longer time in hospitals, which leads to increased charges. HOSP_REGION and ZIPINC_QRTL also play a role in predicting total costs.

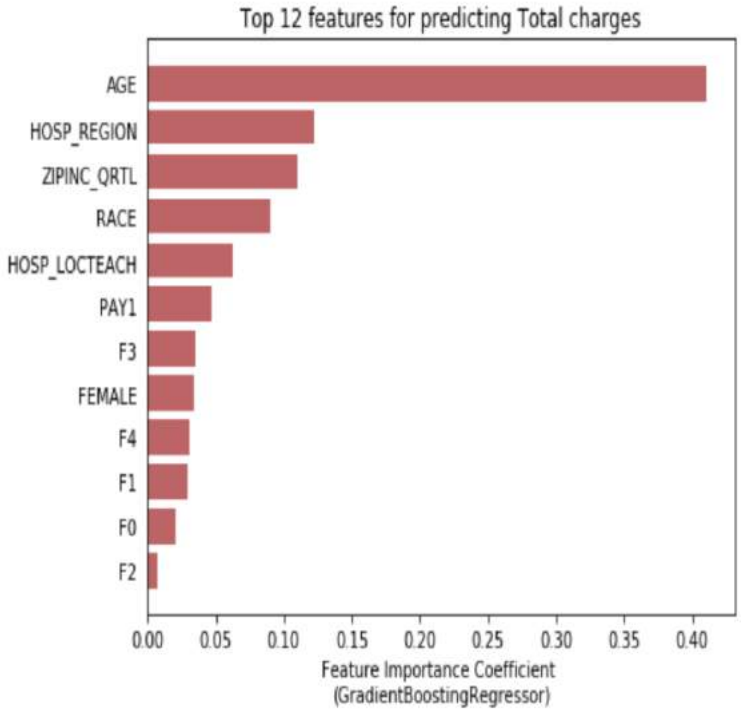

Fig 100

6. Findings

From all the above implementations, we can conclude many factors affect lung cancer and mental illness deaths. There is a slight likely correlation between lung cancer and mental illness. Patients who have lung cancer and having mental illness diagnosis categories F0(Mental disorders due to physiological conditions) and F7(Mental Retardation) show some strong correlation compared to other codes. Heatmaps and Chi-square tests have been performed to find the correlation. Predicted Length of Stay for the patient's suffering from lung cancer and different categories of mental illness. We have implemented various modeling techniques to predict LOS, where Gradient Boosting and Linear Regression stood at the top by giving low error values. Also, predicted total charges that a patient will pay if he gets admitted to the hospital when affected with lung cancer and SMI. What different Socio-Demographic factors affect LOS and Total Charges were also represented. A UI was designed and implemented, which predicts LOS and Total charges when gave particular inputs. This might be helpful for doctors and patients.

7. Summary

After a lot of analysis of data, we discovered many new insights. We found that there is a slight positive correlation between Mental Disorders due to psychological conditions and Death. Likewise, patients with Mental Retardation are slightly correlated to Death. Our findings indicate that patients belonging to age group 60-80 years and patients diagnosed with Mental disorders due to known physiological conditions, Schizophrenia, Mental retardation stayed longer in hospitals and hence paid more charges. Many minor and major challenges were faced while working on this project. In the first problem statement, while finding a correlation between Mental illness and Lung

cancer, we met a potential risk of identifying correlation while grouping people with multiple codes. The most challenging part of this project was filtering the code and feature engineering of the SMI codes into a more usable and interpretable form.

8. Future Work

Here lies the most crucial area for future improvement. As diagnosis has such substantial feature importance, it would be worth evaluating whether an additional sub-division of ICD-10 categories would yield a better prediction model for predicting LOS. This might result in low RMSE and indicating better LOS. In the data provided by HCUP, subdivisions for ZIPINC_QRTL were given in terms of income but not in ZIP codes. Suppose we receive how the categorization of median income is done according to zip codes. In that case, it will be more helpful for the developed UI. In the future, we can predict the risk of unplanned readmissions for this category of patients. Probability of Complication for this category of patients can be predicted.

9. References

Bibliography

American Cancer Society. (n.d.). Retrieved from cancer.org: https://www.cancer.org/about-us/who-we-are.html

Brownlee, J. (2018, June 15). *A Gentle Introduction to the Chi-Squared Test for Machine Learning*. Retrieved from Machine Learning Mastery: https://machinelearningmastery.com/chi-squared-test-for-machine-learning/

codes, D. (n.d.). *hospital los predictor*.

JAMA NETWORK. (2013, February). Retrieved from Jama Physiatry: https://jamanetwork.com/journals/jamapsychiatry/fullarticle/1485447

Services, D. o. (n.d.). *HCUP-US*. Retrieved from HCUP: https://www.hcup-us.ahrq.gov

11. Appendix B

Risks and Mitigations

Risk Name	Description	Probability	Impact	Mitigation
Correlation Between Mental Illness, Lung Cancer and Death	People having multiple codes are grouped together as single category. This could be a potential risk in identifying correlation	Low	Low	Creating multiple columns for groups(Patient's having multiple codes)
Model Accuracy	We have only 5581 records (patients with lung cancer and mental illness) . Less number of records may be one of the reason for poor model accuracy	Low	Low	