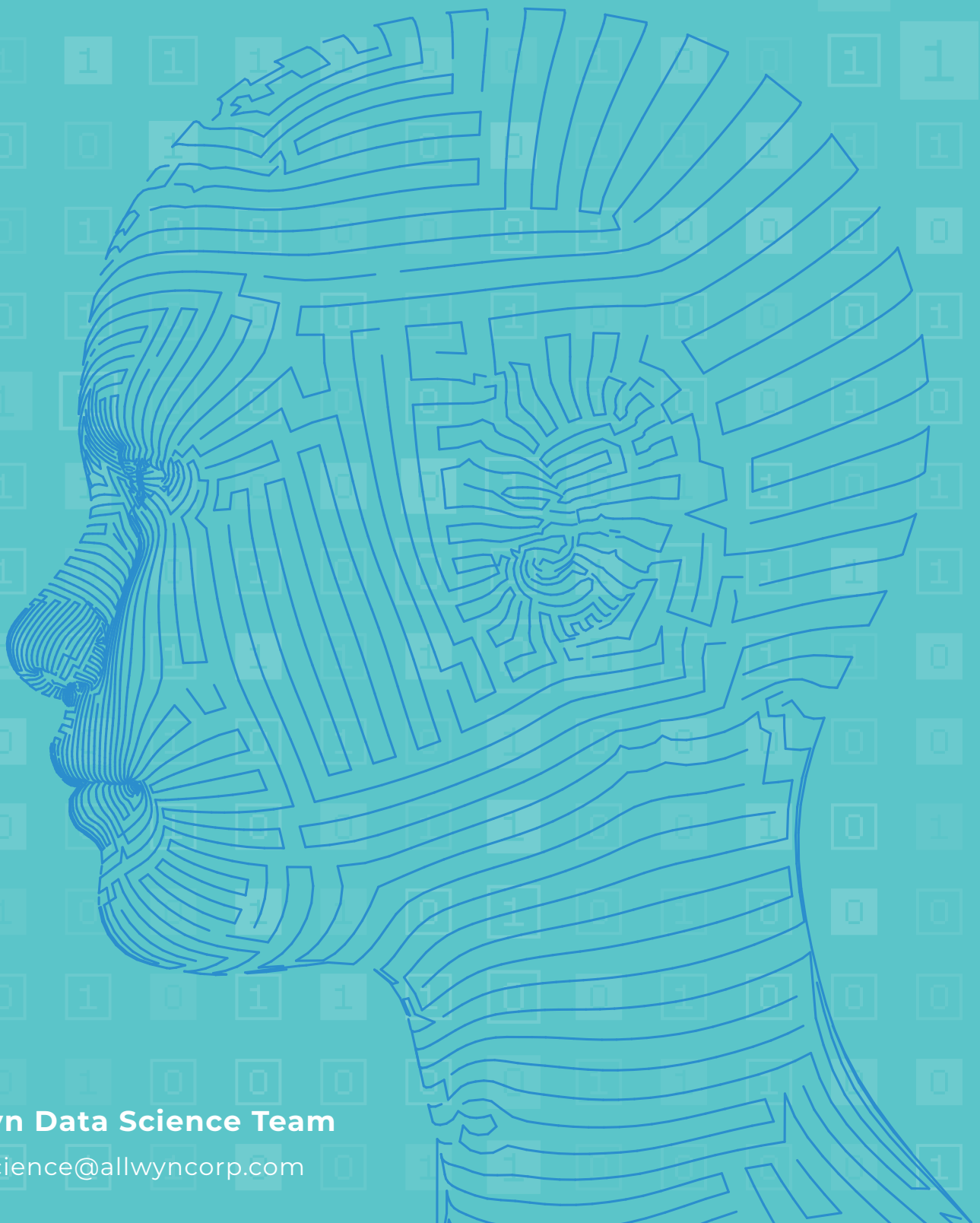
ALLWYN CORPORATION

USING AI TO PREDICT OUTCOMES

for lung cancer patients with severe mental illness



Allwyn Data Science Team

datascience@allwyncorp.com

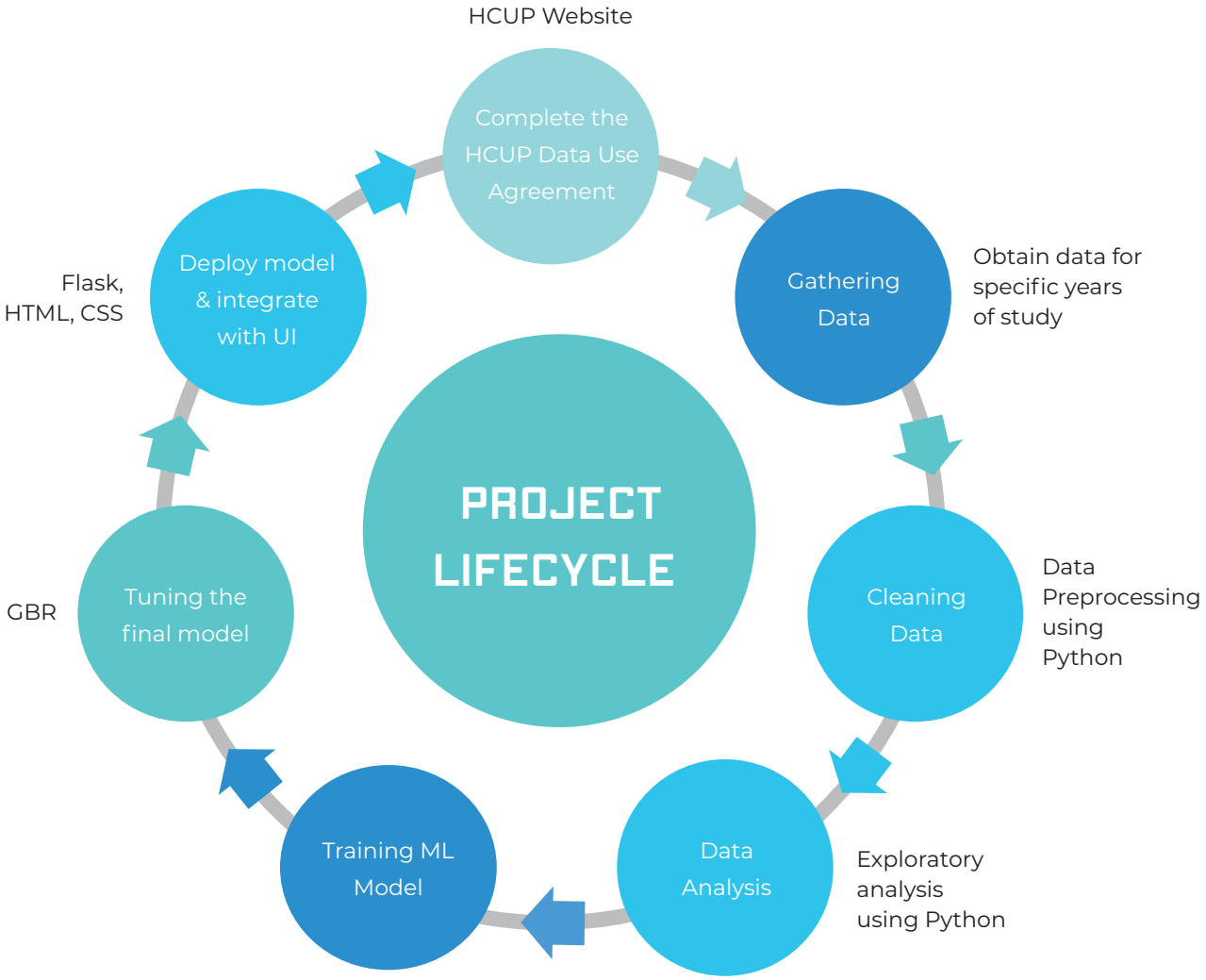
EXECUTIVE SUMMARY

Lung cancer is the number one cause of cancer-related deaths worldwide. Patients with severe mental illness (SMI) are a group who are overrepresented in the lung cancer population. SMI refers to psychological problems, including mood disorders, major depression, schizophrenia, bipolar disorder, and substance abuse disorders, that inhibit a person's ability to engage in functional and occupational activities. Cancer patients diagnosed with SMI may not adhere to treatment plans and may have reduced access to healthcare. Individuals with SMI may have advanced tumor growth at diagnosis due to factors such as limited access to healthcare and healthcare systems. The aggregation of inadequate healthcare and increased risk for somatic disorders in patients with SMI can explain higher mortality rates. Many research papers have indicated that cancer represents a significant proportion of excess mortality for people with mental illness. Mental illness is typically associated with suicide, but much of the excess mortality rates associated with mental illness are due to cardiovascular or respiratory diseases and cancer.

In order to improve outcomes for lung cancer patients with SMI, it is important to study and understand the factors associated with patients who have a mental illness and whether they have a worse case fatality associated with cancer.

In this study, we specifically focused on lung cancer patients who have undergone lobectomy (lung cancer surgery) and analyze if any specific mental illness/psychiatric diagnoses or groups of diagnoses increase perioperative death risk. We also tried to understand the correlation between patients who have undergone lobectomy and its effect on their length of stay in the hospital and the costs associated with it.

PROJECT LIFECYCLE



DATA ANALYSIS:

Data was used from HCUP (Healthcare Cost and Utilization project in the United States) and includes dataset from NIS(National Inpatient sample database), which derives its data from billing data submitted by hospitals statewide across the U.S. This data represents a 20% sample of all the hospitalizations in the U.S. The data used for this study was from NIS 2016 and NIS 2017 data. Datasets contained data related to patient diagnosis and procedure codes in ICD-10-CM/PCS format.

BELOW IS A SNAPSHOT OF THE SAMPLE DATA

	AGE	DRG	FEMALE	LOS	HOSP DIVISION	HOSP NIS	HOSP BEDSIZE	HOSP LOCTEACH	HOSP REGION	TOTC
COUNT	7.159352e+06	7.159694e+06	7.158744e+06	7.159362e+06	7.159694e+06	7.159694e+06	7.159694e+06	7.126364e+06	7.126364e+06	7.127914e-00
MEAN	4.957103e+01	5.464146e+02	5.639609e-01	4.621081e+00	5.071960e+00	5.102384e+04	2.306799e+00	2.597629e+00	2.610482e+00	4.978758e-00
STD	2.738209e+01	2.725937e+02	4.958922e-01	6.914585e+00	2.441306e+00	2.442380e+04	7.819409e-01	6.486434e-01	1.003058e+00	9.615569e-00
MIN	0.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	1.000300e+04	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e-00
25%	2.900000e+01	2.930000e+02	0.000000e+00	2.000000e+00	3.000000e+00	3.033700e+04	2.000000e+00	2.000000e+00	2.000000e+00	1.296900e-00
50%	5.600000e+01	5.920000e+02	1.000000e+00	3.000000e+00	5.000000e+00	5.032400e+04	3.000000e+00	3.000000e+00	3.000000e+00	2.684100e-00
75%	7.200000e+01	7.940000e+02	1.000000e+00	5.000000e+00	7.000000e+00	7.047600e+04	3.000000e+00	3.000000e+00	3.000000e+00	5.456800e-00
MAX	9.000000e+01	9.990000e+02	1.000000e+00	3.650000e+02	9.000000e+00	9.053800e+04	3.000000e+00	3.000000e+00	4.000000e+00	9.999999e-00

The initial data size, which includes all lobectomy cases for 2016 and 2017, was 13,892 cases in the HCUP NIS database. Filtering these cases for SMI's resulted in a dataset that is 5581 cases. The following groups of diagnoses codes were considered in our study.

Mental Illness Diagnoses Code Groups

Section F40-F48	Anxiety, dissociative, stress-related, somatoform and other nonpsychotic mental disorders (F40-F48)
Section F60-F69	Disorders of adult personality and behavior (F60-F69)
Section F50-F59	Behavioral syndromes associated with psychological disturbances & physical factors (F50-F59)
Section F80-F89	Pervasive and specific development disorders (F80-F89)
Section F30-F39	Mood (affective) disorders (F30-F39)
Section F20-F29	Schizophrenia, schizotypal, delusional,and other non mood psychotic disorders (F20-F29)
Section F90-F98	Behavioral and emotional disorders with onset usually occuring in childhood & adolescence (F90-F98)
Section F10-F19	Mental and behavioral disorders due to psychoative substance use (F10-F19)
Section F01-F09	Mental disorders due to known physiological conditions (F01-F09)

At Allwyn, we use big data techniques to extract, transform and load the data to conclude with a meaningful dataset that is quality controlled as well as engineered for any missing values, outliers, and grouping of mental illness codes. We also ensure that we balance the datasets so that the results are not skewed.

We explored hundreds of data elements along with our subject matter expert to understand the significance of diagnoses codes, procedure codes, or discharge information. We focused on the categorical variables and performed additional research based on our SME's inputs as well as obtaining feedback from HCUP.

Our core data file included important data elements such as patient's details such as age, race, and urban/rural location, hospital data like date of admission, hospital location, etc., and important medical information such as procedure codes and Diagnosis codes. Some of the important features of the dataset used for this study are included below:

Mental Illness & Lung Cancer - Dataset Description

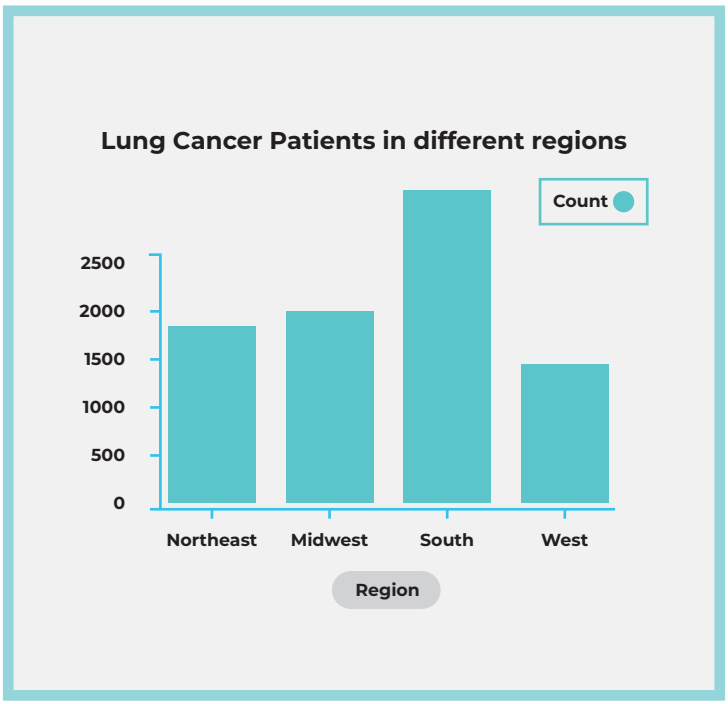
Data Description

ATTRIBUTE	DESCRIPTION OF CODE
AGE	Age in years coded are 0-124
FEMALE	Indiates gender for NIS begining in 1998: (0) male, (1) female
I10_DX-I10_DX40	ICD-10-CM diagnoses, principal and secondary, with external cause of morbidity codes at the end of the array
I10_NDX	Number of ICD-10-CM diagnoses coded on the record
DIED	Indicates in-hospital death: (0) did not die during hospitalization, (1) died during hospitalization
YEAR	Calendar year
DRG	DRG in use on discharge date (based on ICD-10-CM/PCS Codes)
I10_PR1-I10_PR25	ICD-10-PCS procedures, principal and secondary
I10_NRP	Number of ICD-10-PCS procedures coded on the record
PRDAY 1	Number of days from admission to principal procedure

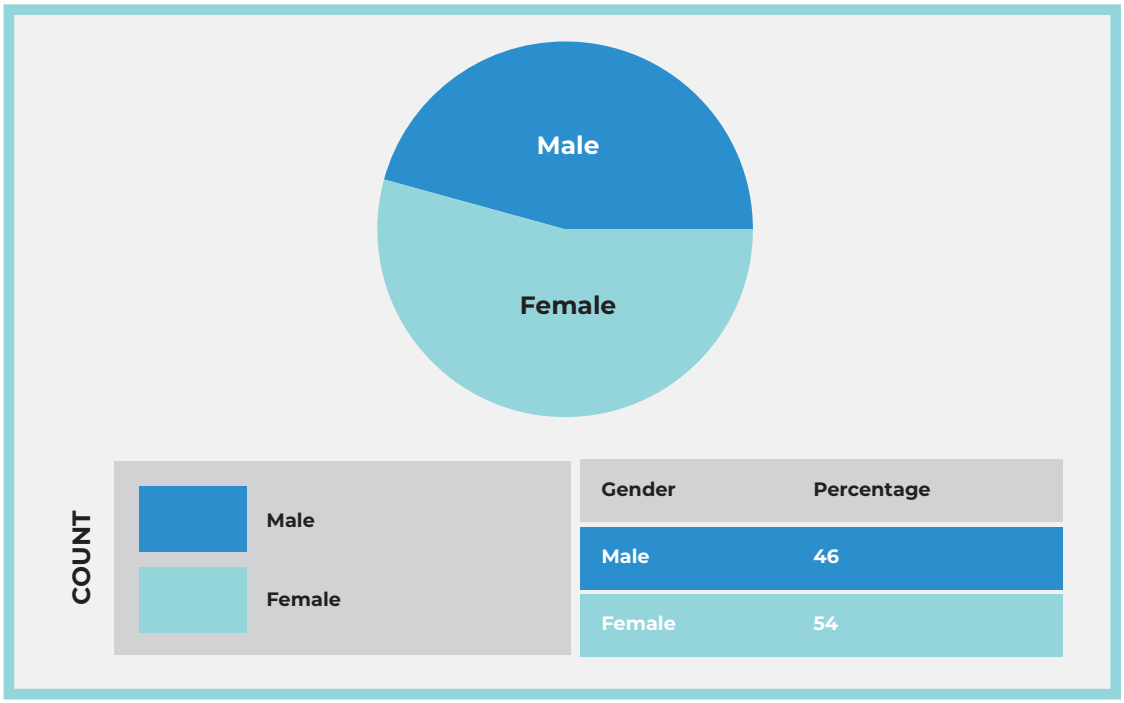
ATTRIBUTE	DESCRIPTION OF CODE
TOTCHG	Total charges, edited
LOS	Length of stay, edited
YEAR	Calendar year
KEY_NIS	Unique record number for file beginning in 2012 links the Core File to other discharge-level NIS files.
DISCWT	Discharge weight on Core File and Hospital File for NIS beginning in 1998.
HOSP_NIS	NIS hospital number (links to Hospital File; does not link to previous years)
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
KEY_NIS	Unique record number for file beginning in 2012 links the Core File to other discharge-level NIS files.
DISCWT	Discharge weight was used in the NIS beginning in 1998.
APRDRG	All Patient Refined DRG
HOSP_NIS	NIS hospital number (links to Hospital File; does not link to previous years)
KEY_NIS	Unique record number for file beginning in 2012
HOSP_BEDSIZE	Bed size of the hospital (STRATA): (1) small, (2) medium, (3) large

DATA DISTRIBUTION

For any machine learning algorithm to be designed, it is important to understand the variability of the data and skewness, as well as the assumptions that we can make to build machine learning models. Here are some key statistical distribution models of the dataset we used for our study:

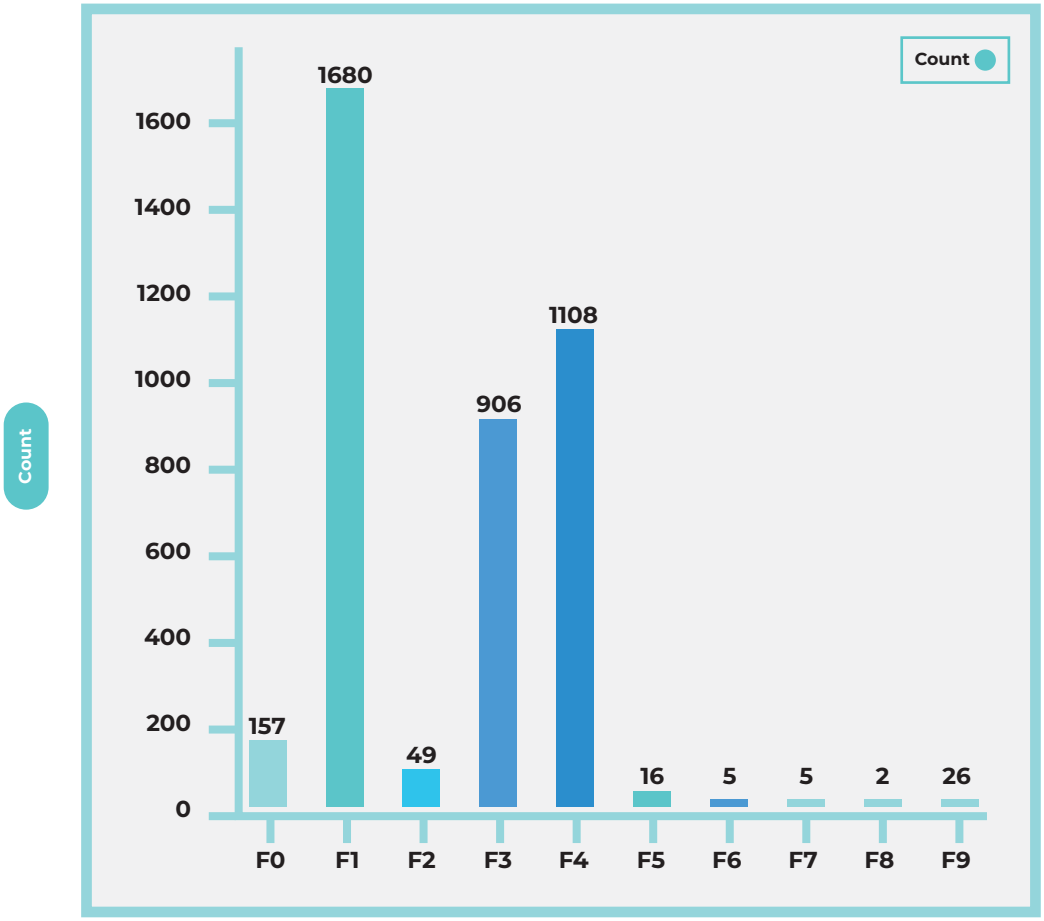


PERCENTAGE OF MALE AND FEMALE LUNG CANCER PATIENTS



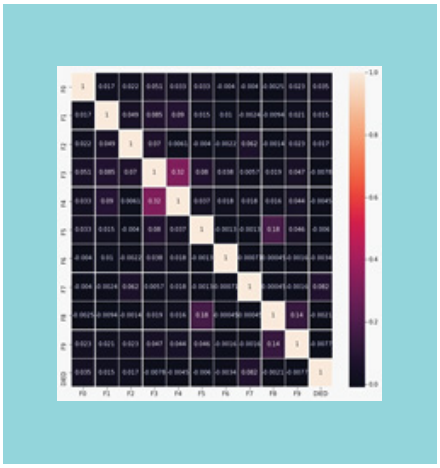
NUMBER OF PATIENTS WITH DIFFERENT MENTAL ILLNESS

By grouping of the mental illness diagnoses codes after filtering Lobectomy procedure codes, we observed a majority of data to be falling under F1, which refers to F10-F19.

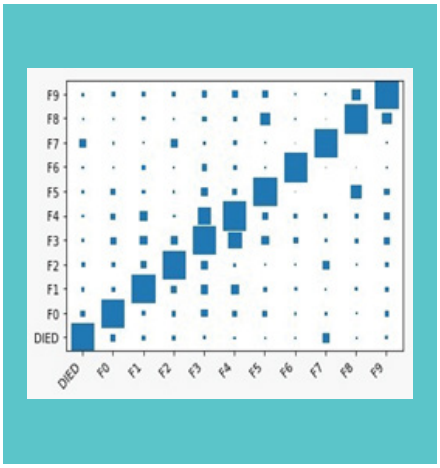


Section F40-F48	Anxiety, dissociative, stress-related, samatoform & nonpsychotic mental discorders (F40-F48)
Section F60-F69	Disorder of adult personality and behaviour (F60-F69)
Section F50-F59	Behavioral syndromes associated with physiological disturbances and physical factors (F50-F59)
Section F80-F89	Pervasive and specific development disorders (F80-F-89)
Section F30-F39	Mood (affective) disorders (F-30-F39)
Section F20-F29	Schizophrenia, schizotypal, delusional, and other non-mood psychotic disorders (F20-F29)
Section F90-F98	Behavioral and emotional disorders with onset usually occuring in childhood & adolescence (F90-F98)
Section F10-F19	Mental & behavioral disorders due to psychaotive substance use (F10-F19)
Section F01-F09	Mental disorders due to known physiological conditions (F01-F09)

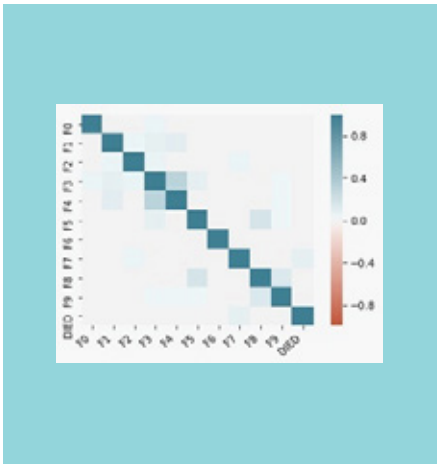
CO-RELATION BETWEEN MENTAL ILLNESS AND LUNG CANCER AND DEATH



Coefficient Representation



Scatter Plot Representation



Correlation Matrix

From the above plots, there is a slight correlation between F0 (Mental disorders due to known physiological conditions), F7(Mental retardation) with death.

DATA DISTRIBUTION OF LENGTH OF STAY IN THE HOSPITAL FOR PATIENTS WHO HAVE UNDERGONE LOBECTOMY AND HAVE A MENTAL ILLNESS

We observed that patients with mental illness codes F0, F7, and F8 stayed longer. Patients with F0, F2 paid more charges in the hospital than other mental illness groups.

F0 - Mental disorders due to known physiological conditions

F1 - Mental and behavioral disorders due to psychoactive substance use

F2 - Schizophrenia, schizotypal, delusional, and other Non-mood psychotic disorders

F3 - Mood [affective] disorders

F4 - Anxiety, dissociative, stress-related, somatoform and other nonpsychotic mental disorders

F5 - Behavioral syndromes associated with physiological Disturbances and physical factors

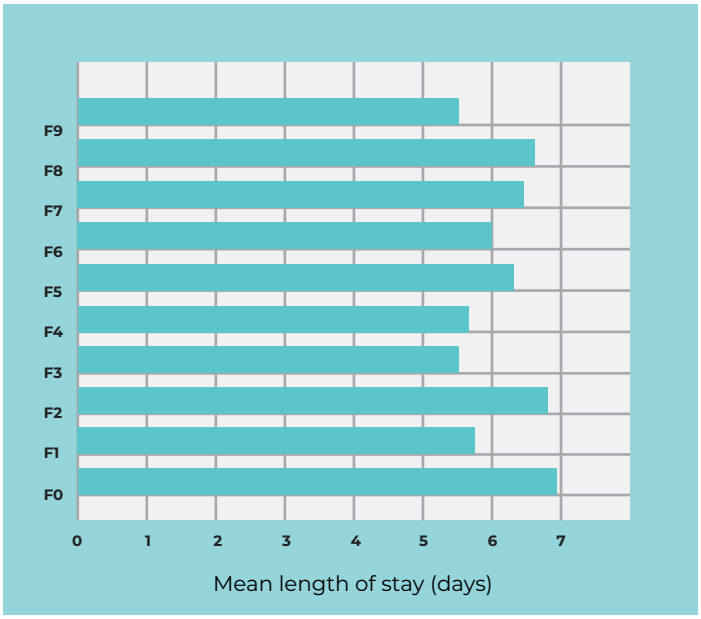
F6 - Disorders of adult personality and behavior

F7 - Mental retardation

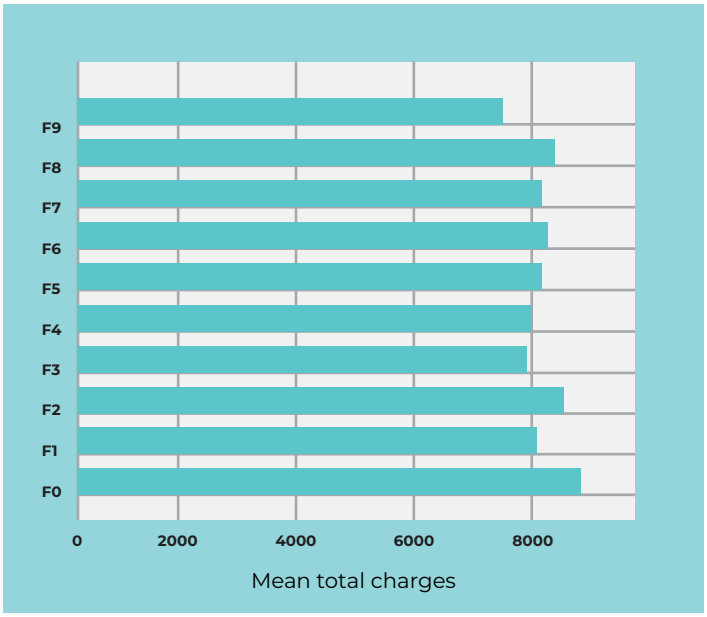
F8 - Pervasive and specific developmental disorders

F9 - Behavioral and emotional disorders with onset usually occurring in childhood and adolescence

Comparison of Diagnoses



Comparison of Diagnoses codes



MODELING AND ALGORITHM

In order for machine learning algorithms to provide less degree of variability and a higher level of accuracy, the data is generally split into two different segments — training and testing. The algorithm is trained on a partial set of data called training, and once satisfied with the results, the algorithm is then run on the testing set. The performance of the model is measured in the training data across the various types of machine learning models.

PROBLEM STATEMENT

For determining the correlation between lung cancer patients who have undergone lobectomy and have a mental illness, the team developed multiple machine learning models. For this study, we split the data into 80% training data (4464 sample data) and 20% test data (1117 sample data).

We divided this problem statement into two areas of evaluation:

- Predicting LOS of a patient with both lung cancer and mental illness using only Diagnosis codes.
- Predicting LOS of a patient with both lung cancer and mental illness using both Diagnosis codes and Socio-demographic features.

The following algorithms were then developed:

- SGDRegressor
- GradientBoostingRegressor
- LinearRegression
- KNeighborsRegressor
- RandomForestRegressor
- SVR
- TensorFlow

PROBLEM IA: PREDICTING LOS OF A PATIENT WITH BOTH LUNG CANCER AND MENTAL ILLNESS USING ONLY DIAGNOSIS CODES.

Here are some examples of the algorithms developed: Below is a **Gradient Boosting Regression Algorithm :**

Prediction Model days 1.6898821756888593

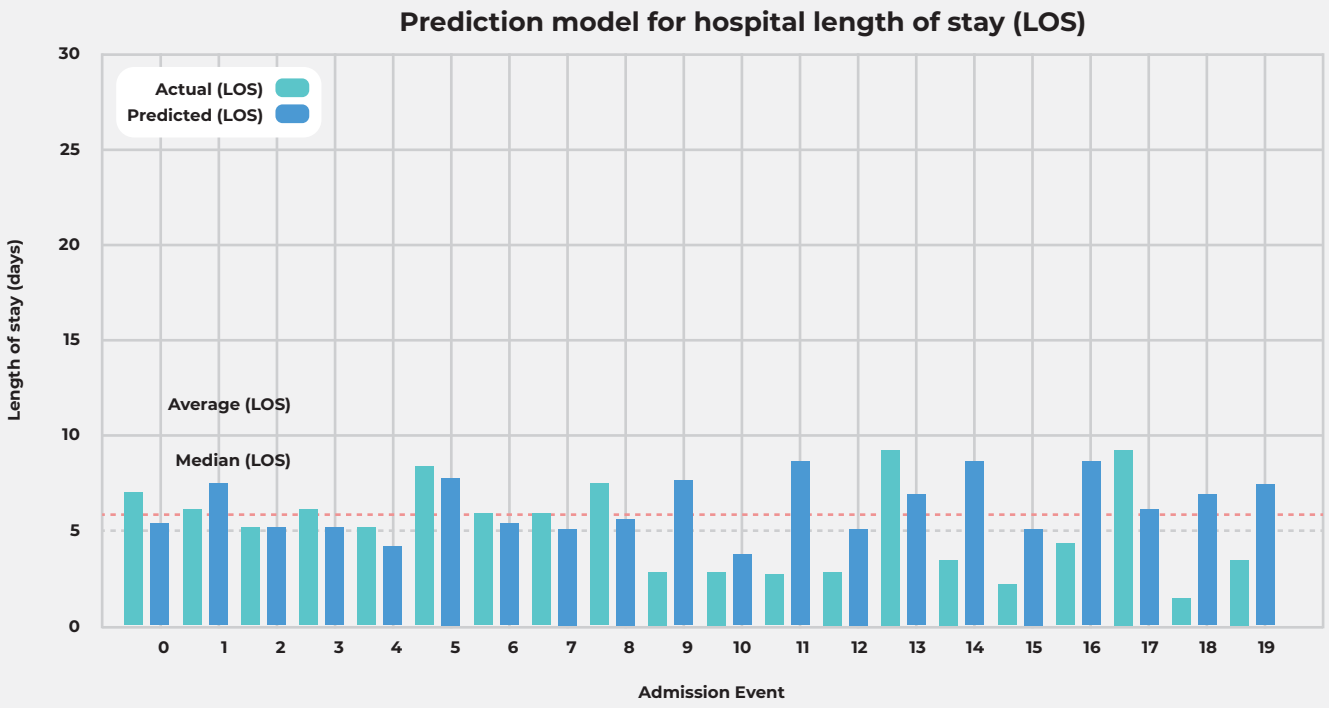
Median Model days 2.2820053715308863

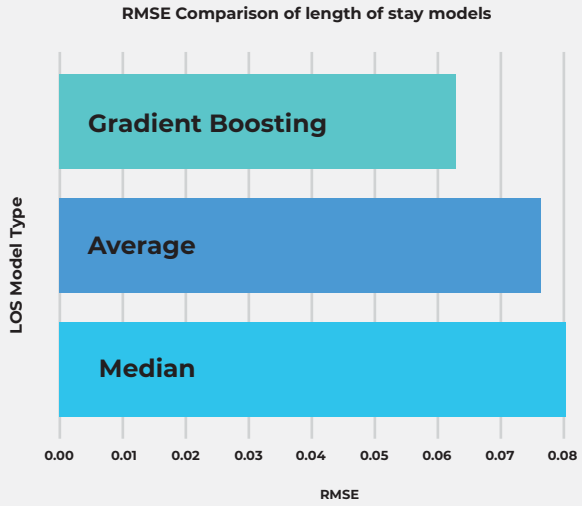
Average Model days 2.267122095573982

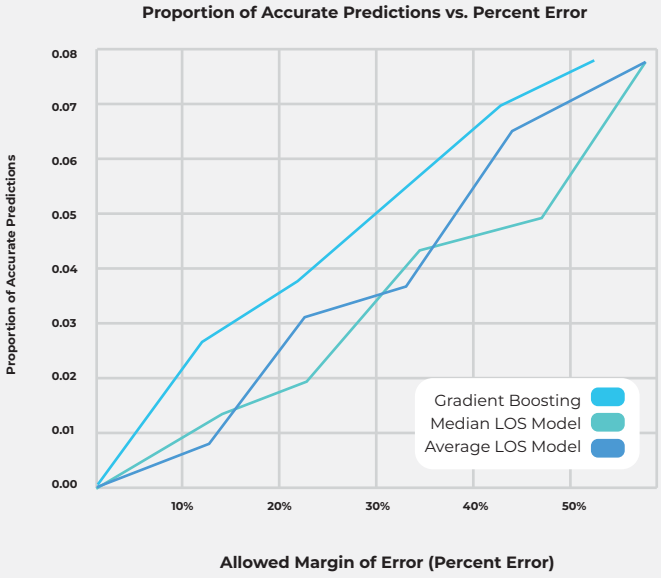
Prediction Model RMS 0.06232652015934372

Median Model RMS 0.7993881552828291

Average Model RMS 0.07624966411136928







Training set has 4464 samples

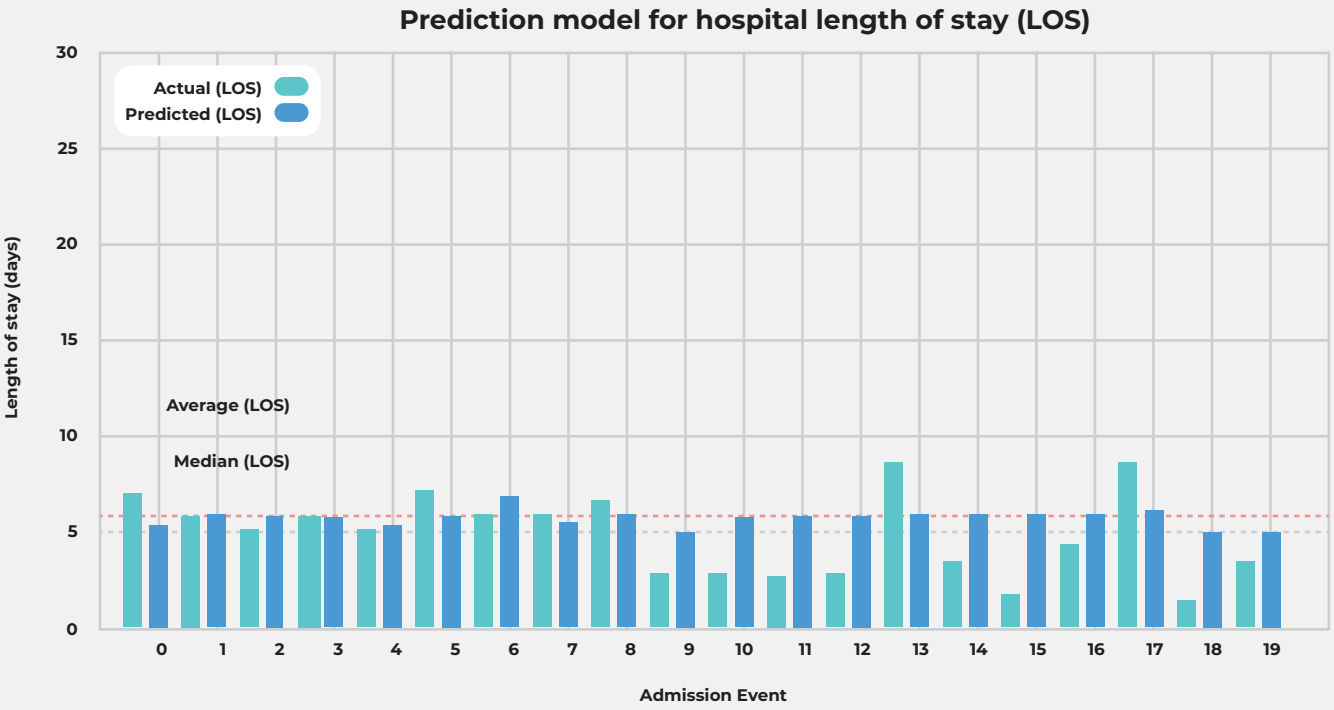
Testing set has 1117 samples

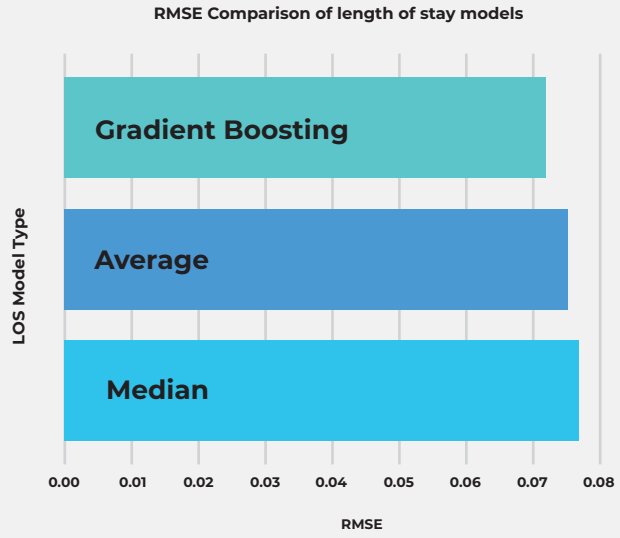
R2 score is : 0.331133

MAE score is : 1.689882

MSE score is : 4.339093

Max error score is : 5.794150





Training set has 4464 samples

Testing set has 1117 samples

RMSE score is: 2.440360

R2 score is : 0.019399

MAE score is : 2.120314

MSE score is : 5.955358

Max error score is : 5.324139

Prediction Model days 1.6937242187594366

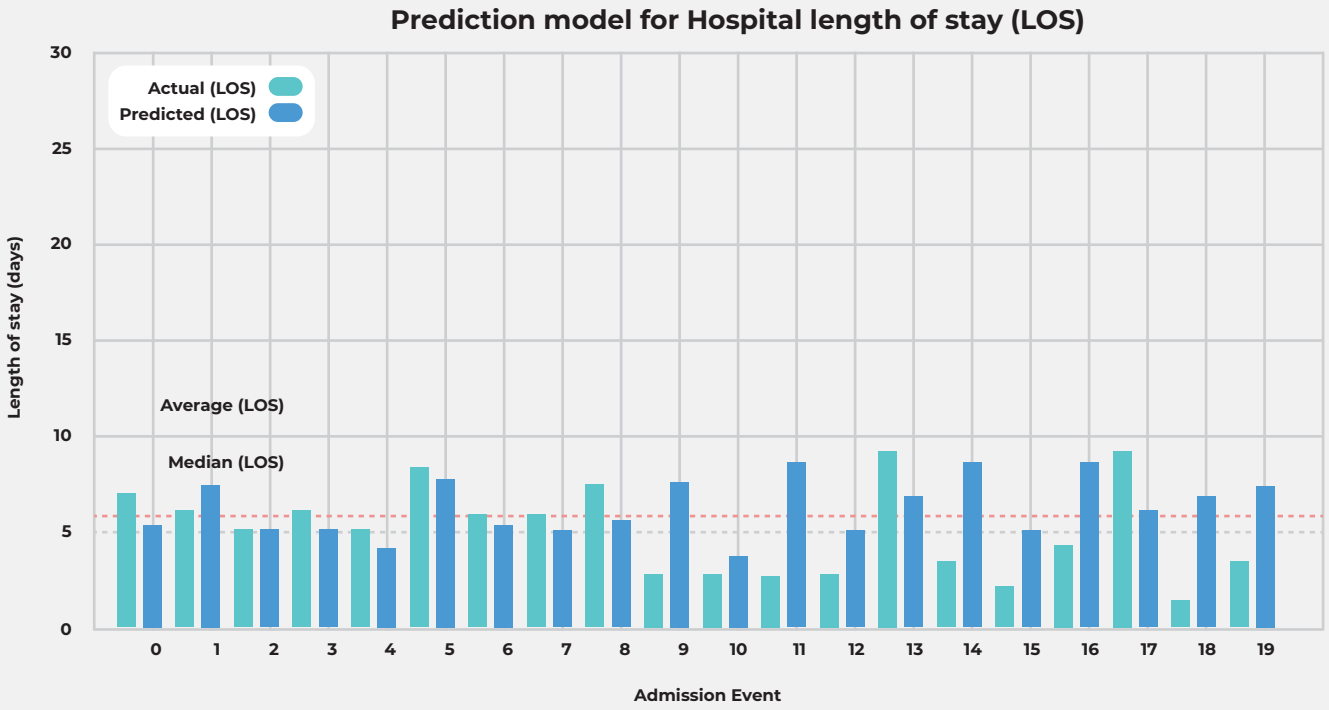
Median Model days 2.2820053715308863

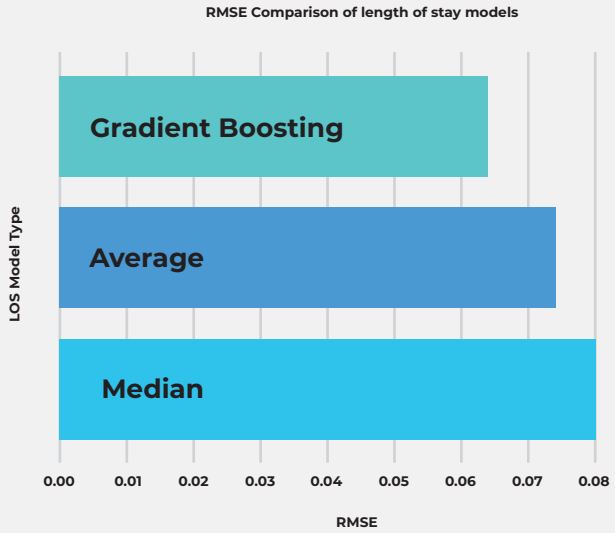
Average Model days 2.267122095573982

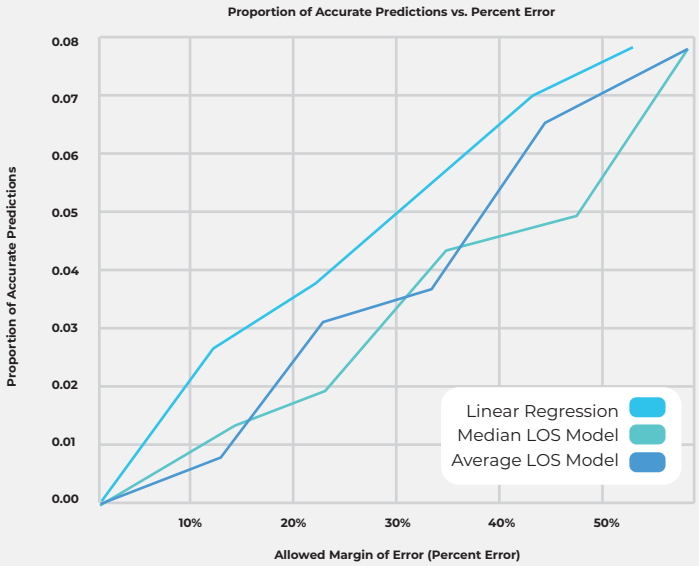
Prediction Model RMS 0.06208706861923202

Median Model RMS 0.7993881552828291

Average Model RMS 0.07624966411136928







Training set has 4464 samples

Testing set has 1117 samples

R2 score is : 0.336262

MAE score is : 1.693724

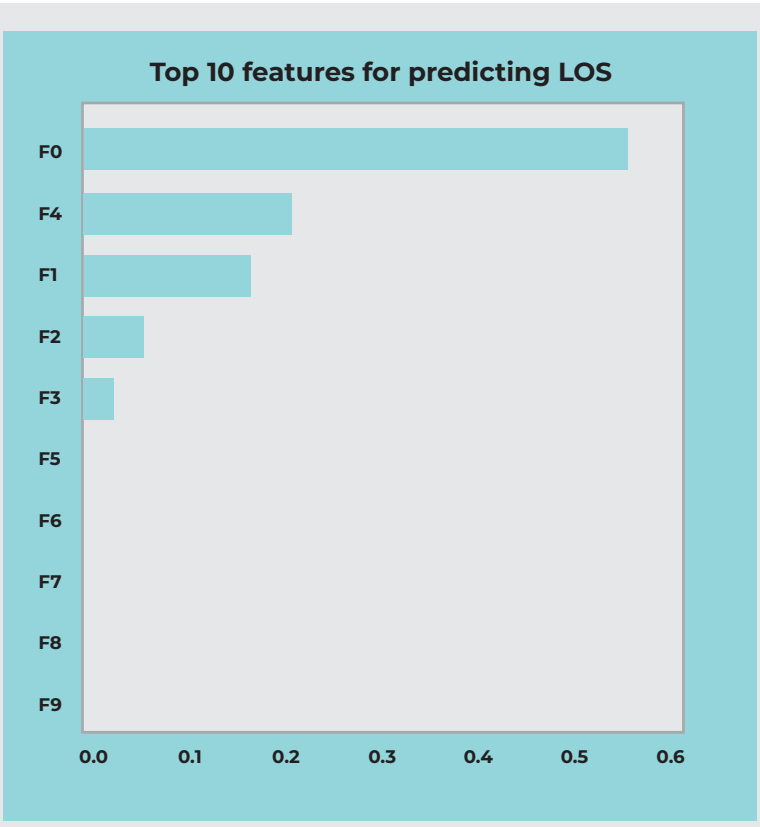
MSE score is : 4.305816

Max error score is : 5.778020

Comparing the model performance of the different machine learning models showed that GBR had the minimum MAE.

Model	R2	RMSE	MAE	MSE	MAX ERROR
GBR	0.010	2.440	2.120	5.955	5.324
SVR	-0.015	2.567	2.218	6.590	5.900
SGD	0.0157	2.524	2.230	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

GBR was then used to perform cross-validation with KFold Split by 4, and a feature importance plot for GBR was developed.



Feature Importance Coefficient
(Gradient boosting regressor)

F0 - Mental disorders due to known physiological conditions

F1 - Mental and behavioral disorders due to psychoactive substance use

F2 - Schizophrenia, schizotypal, delusional, and other non-mood psychotic disorders

F3 - Mood (affective) disorders

F4 - Anxiety, dissociative, stress-related, somatoform and other nonpsychotic mental disorders

F5 - Behavioral syndromes associated with physiological disturbances and physical factors

F6 - Disorders of adult personality and behavior

F7 - Mental retardation

F8 - Pervasive and specific developmental disorders

F9 - Behavioral and emotional disorders with onset usually occurring in childhood and adolescence

In conclusion, , it is clear that F0(Mental disorders due to physiological conditions) is the top feature for predicting LOS, followed by F4(Anxiety, dissociative, stress-related, and other nonpsychotic mental disorders) and F1(Mental and behavioral disorders due to psychoactive substance use).

Problem 1B: Predicting LOS of a patient with both lung cancer and mental illness using both Diagnoses codes and Socio-demographics. Machine learning models were developed with both mental illness diagnosis and socio-demographics such as age, gender, and income quartiles. Here is the performance of the various machine learning models:

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

From the above results, GBR performed well with low RMSE, MAE, MSE, and Max error.

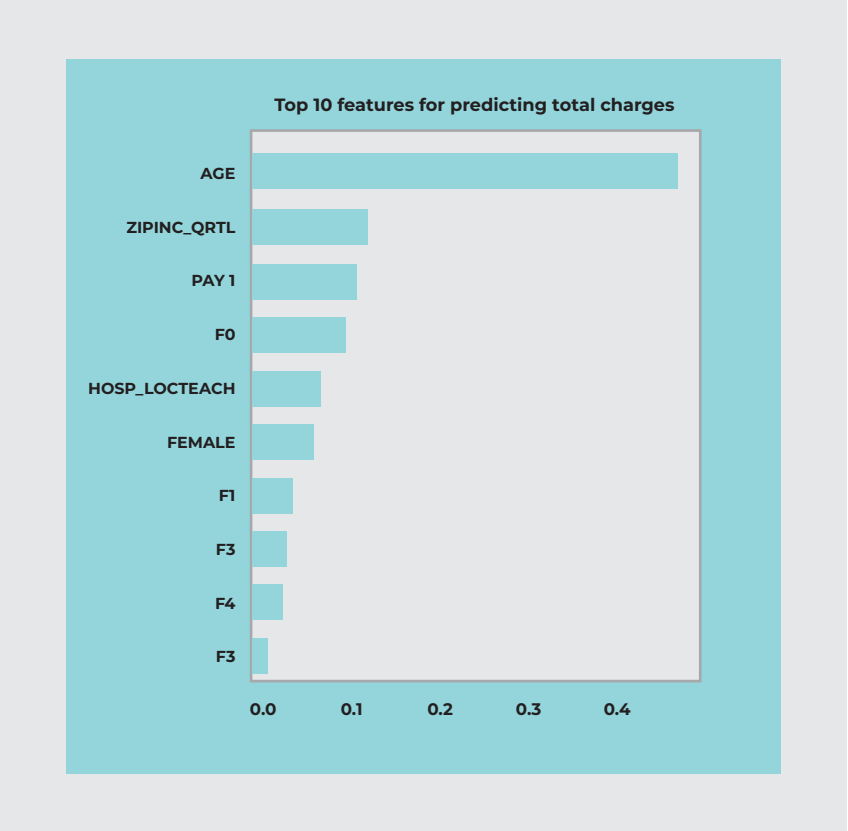
F0 - Mental disorders due to known physiological conditions

F1 - Mental and behavioral disorders due to psychoactive substance use

F2 - Schizophrenia, schizotypal, delusional, and other non-mood psychotic disorders

F3 - Mood (affective) disorders

F4 - Anxiety, dissociative, stress-related, somatoform and other nonpsychotic mental disorders

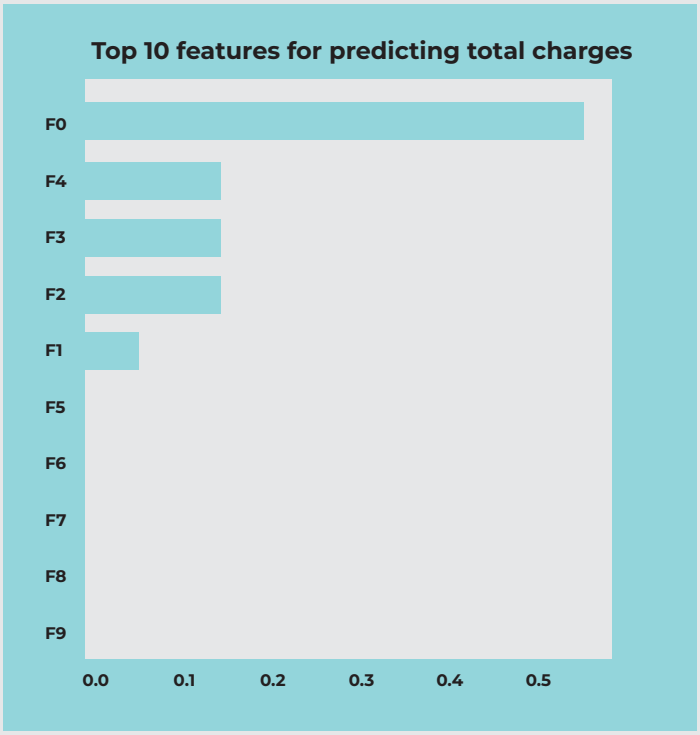


In conclusion, as per the feature importance plot, it is clear that AGE is the most important factor for predicting length of stay, followed by ZIPINC_QRTL(Median Income quartiles) and PAY1(Primary Payer information). It is understandable that Aged people will stay longer in hospitals.

Problem 3a: : Predicting the total charges for a patient with both lung cancer and mental illness using Diagnosis codes.

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

From the above results, both Gradient Boosting and Linear regression model performed well with low RMSE, MAE, MSE, and Max error.



F0 - Mental disorders due to known physiological conditions

F1 - Mental and behavioral disorders due to psychoactive substance use

F2 - Schizophrenia, schizotypal, delusional, and other non-mood psychotic disorders

F3 - Mood (affective) disorders

F4 - Anxiety, dissociative, stress-related, somatoform and other nonpsychotic mental disorders

F5 - Behavioral syndromes associated with physiological disturbances and physical factors

F6 - Disorders of adult personality and behavior

F7 - Mental retardation

F8 - Pervasive and specific developmental disorders

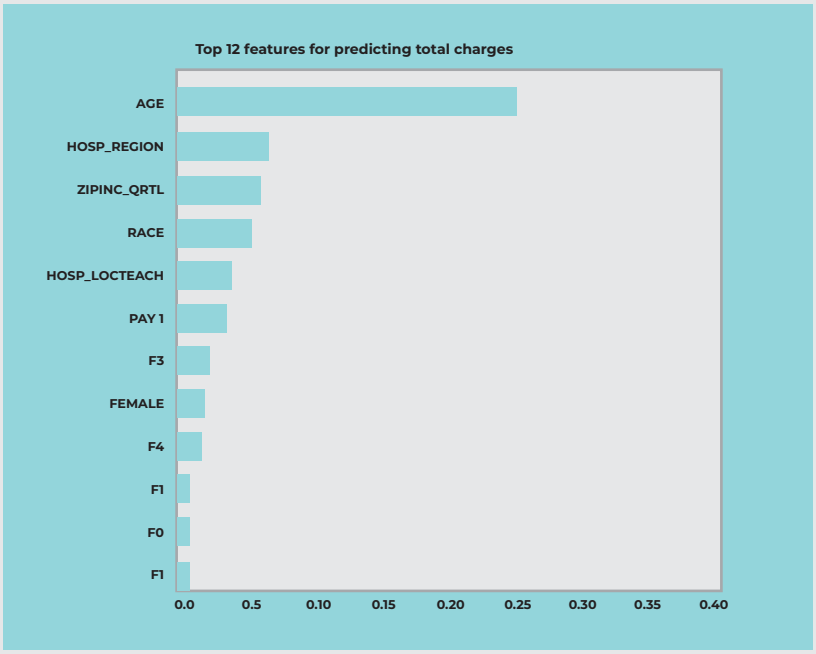
F9 - Behavioral and emotional disorders with onset usually occurring in childhood and adolescence

As per the feature importance plot, it is clear that F0 (Mental disorders due to physiological conditions) is the top feature for predicting TOTCHG, followed by F4 and F2. It is understandable that patients with mental disorders are affected much and stayed for a longer period in hospital, so they have to pay more charges.

Problem3b: Predicting the total charges for a patient with both lung cancer and mental illness using both Diagnoses codes and Socio-demographics such as Age, Sex, Race, Median household income for patient's Zip code), expected primary payer, and hosp_locteach (rural, urban nonteaching, urban teaching) and HOSP_REGION.

Model	R2	RMSE	MAE	MSE	MAX ERROR
GBR	0.027	0.386	0.315	0.149	2.570
SVR	-0.068	0.405	0.302	0.164	2.713
Linear Regression	0.022	0.387	0.318	0.150	2.593
Random Forest	-0.174	0.425	0.336	0.180	2.681

From the above results, we find that Gradient Boosting Regression performed well with low RMSE, MAE, MSE and Max error values.



F0 - Mental disorders due to known physiological conditions

F1 - Mental and behavioral disorders due to psychoactive substance use

F2 - Schizophrenia, schizotypal, delusional, and other non-mood psychotic disorders

F3 - Mood (affective) disorders

F4 - Anxiety, dissociative, stress-related, somatoform and other nonpsychotic mental disorders

F5 - Behavioral syndromes associated with physiological disturbances and physical factors

F6 - Disorders of adult personality and behavior

F7 - Mental retardation

F8 - Pervasive and specific developmental disorders

F9 - Behavioral and emotional disorders with onset usually occurring in childhood and adolescence

As per the feature importance, it is clear that AGE is the top feature. HOSP_REGION and ZIPINC_QRTL also play a role in predicting total charges.

LIMITATIONS

As with any machine learning/AI project, more data can yield better results. We have used two years of data for this exploratory study and hence were limited to the results we have achieved.

Other limitations include not gaining a high level of accuracy for parts 2 and 3 regarding the relationship between the length of stay in the hospital with regards to socio-economic parameters.

CONCLUSION

Here are the results for all three problems studied for patients who have undergone lobectomy (lung cancer surgery) and have mental illnesses.

In the case of patients with mental illness who have undergone lobectomy, analysis of the length of stay at the hospital revealed that F0 (Mental disorders due to physiological conditions) is the top feature affecting the longer length of stay, followed by F4 (Anxiety, dissociative, stress-related and other nonpsychotic mental disorders) and F1(Mental and behavioral disorders due to psychoactive substance use).

When considering the various socio-demographic factors, we find that AGE is the most important factor affecting both length and cost of the stays in the case of Lobectomy patients with SMI.

This study was performed in collaboration with Dr. James Baldo and Dr. Isaac Gang of the DAEN program of the Volgeneau School of Engineering at George Mason University, Fairfax VA. Several students of the DAEN program contributed to the study.

ABOUT ALLWYN CORPORATION

Allwyn Corporation (www.allwyncorp.com) is a forward thinking, innovative software solutions company, headquartered in the Metropolitan Washington DC area. Allwyn was founded in 2003 with a mission to help organizations address complex technology challenges by providing industry-leading tools, technologies, seasoned professionals, and proven methodologies. We are proud to be certified for ISO 9001 (Quality), ISO 27001 (Security), and ISO 20000 (Service Delivery).
With a team of ~200 professionals, Allwyn delivers high-quality services to a wide range of clients in the public and private sector.

Allwyn has been providing leading-edge IT professional services to various government agencies through the GSA MAS Schedule. We are also on the FAA eFAST, GSA OASIS+, and GSA STARS III contract vehicles.

Allwyn has experience with implementing Artificial Intelligence and Machine Learning solutions and Modernizing Applications using Low Code Technologies. Our relationships with AWS, Appian, ServiceNow, Microsoft, Databricks, Informatica, Salesforce, etc. strengthen our ability to support our customers in their Digital Transformation journey. We are already supporting several of our customers in the public as well as commercial sector with their cloud adoption strategies and Artificial Intelligence and Machine Learning implementations. For additional information on Allwyn's full range of services, please visit our website at www.allwyncorp.com.