Question To Be Answered

1. What are the top causes for a patient being readmitted to the hospital?

1. What if-then rule logic in the Decision Tree increases the probability of a patient being readmitted? Decreases the probability?

The What if-then logic in the decision tree with the largest significance that increases the probability that a patient is readmitted is *Standard Orders.* This creates are largest split in our tree and those without *Standard Orders* have a 34.01% chance of being readmitted compare to those with *Standard Orders* who have a probability of 15.83% of being readmitted. Patients who were in the ICU for less than 17 days had huge increase of being readmitted with a percentage of 92.01%. Also, patients who had the diagnosis of congestion heart failure saw a readmittance rate of 63.04%.

2. What is the "quickest" path to achieve the highest probability of a patient being readmitted?

The find the "quickest" path to achieving the highest probability of a patient being readmitted based on the decision tree model goes as follows

- The *Standard_orders_Used* variable is Yes
- For *Procedure_Long_desc* it is Biopsy Bone Marrow
- The next *Procedure_Long_desc* is Pericardiocentesis

This is the only path in the decision tree where readmittance is 100% in 3 tree leaves.

Introduction

Our analytics team was given four datasets from a large hospital system on various data about patients, hospitals, and geography. The hospital system is having a problem with patients being readmitted back into their system after being discharged and is looking for insights on what factors could be causing this. Throughout this process, the analytics team will conduct Explanatory data analysis to get a deeper insight into the data. From those insights, multiple supervised machine learning algorithms such as logistic regression and decision trees will be used to predict the probability of a patient being readmitted.

Problem Statement

The hospital is looking for a way to find the probability of a patient of being readmitted back into their system after being discharged. With the passage of the Affordable Care Act in 2012 came with the

Hospital Readmission Reduction Program (HRRP). This program financially penalized hospital with a higher than expected risk-standardized 30-day readmission rates. This penalty will continue to grow with every passing year so reducing patient readmittance is a priority. With consumers becoming more conscious of their purchasing decision hospital could lose business to other hospital system who have a much lower readmittance rate due to the quality of care.

Data and Analysis

Before we begin any analysis, we should understand the data that we are working with. We were given four datasets that have information about the demographics of the patient, the condition of the patient in the hospital and information about the initial hospital stay. There also a dataset that has information about the geographical location of the patient.

From this data, we will have to do data preprocessing to ensure the quality and integrity of the data. With the data finally clean and joined together, analysis can be done. Some of the modeling techniques that will be used are logistic regressions and decision trees.

To gain further insight we will use visual analysis on the data using Tableau. This can help visualize our data and create a story for the insights that we found when exploring the datasets.

Findings/Recommendations

When looking at the decision tree model based on the variable used after *Standard_Order_Used* we found that *Procedure_Long_Desc* and *Doctor* had the largest influence on increasing and decreasing the readmittance rate of the patients. Patients who had the procedure of a biopsy of the bone marrow had a significant increase in being readmitted back into the hospital.

While patients go into the hospital for many reasons we found there was one path in particular in the decision tree that dominated the rest. We found that 62203 patients who enter the hospital had these following variables:

- Standard Order Used (Y)
- Procedure_Long_Desc (Diagnostic Ultrasound
- Doctor (287656, 319038, 235415…)
- Diagnosis_Long_Desc (Acute on Chronic Diastolic)
- Disc_Nurse_Id (370015, 15006, 270011….)

We also found that in general patients who didn't use a standard order were much more likely to be readmitted back into the hospital. The percentage of them being readmitted was 34.01% compared to the dataset average of 19.42%.

Patients who didn't have standard order and had the diagnosis of, Chronic Airway Obstruction, Acute Combined Systolic Heart Failure or Acute on Chronic Heart Failure had a significantly larger probability of being readmitted back into the hospital compared to those who didn't.

Limitations and Next Steps

Some of the limitations that could affect our modeling is the amount of data that we have received. While 160,000 record may seem like a lot to some, having a larger data set can allow the model to learn a lot better. There could be also important variables that have strong explanatory value for our target variable that has been left out of the dataset. The next steps for improving our analysis is to go out and gather more data. Through a limitation that can occur from this is that if we have got too much data we might not have enough computing power to make a proper analysis without investing more capital in improving our assets.

Appendix A: Technical Write-Up

Adjusting Variables and Metadata

Similar to the logistic regression model, the variables of our dataset had to be adjusted in order to fit our decision models needs. For the decision tree, all the roles for the variables were kept the same as our logistic regression model except that we rejected *Order_Total_charge*. This was done based on our data science team believes that it has an overwhelming influence on our model, leaving out possible important variables from expressing themselves. For the levels of the variables, they also were kept the same for the exception of the *Doctor* and *Nurse Discharge ID.*

Data Partitioning Node

Just like our logistic regression model we want to split the data that we are using into two separate datasets. We will create two datasets, one for training our decision tree model and the other will be using to validate our decision tree model. The main dataset was split 70/30 with 70% of the data randomly selected in our training dataset and the remaining 30% in our validation dataset.

Decision Tree Node: Properties

The decision tree node has many properties that can be used to change how the model works. Our team decided to keep most of the default properties. We left the default property for “Maximum Branch” and “Maximum Depth” 2 and 6 respectively. Our reasoning for this is to keep a simple interpretable model with the branch only splitting between two rules and ensuring our model is not overfitting with a large depth. The “Missing Values” property was left as it default which was “Use in Search”. Since we decision trees naturally deal with missing data in its rulemaking we left this as the default as we are not using the impute node. With the significance level, it was set at 0.2. Our data team reasoning for this value was that we are working with a decently large dataset, so variables can better express as they do naturally. For the “Leaf Size” we changed the default of 5 to 30. The reasoning for this change is the central limit theorem where a sample must have a large number of records. As long as it is greater than or equal to 30 it will give us the criteria of having a normal distribution which will make our sample variability smaller. This will increase the predictability of our model while also reducing the chance of overfitting our model.

Decision Tree Node: The Tree it and leaves

Once the metadata has been changed, our data partitioning, and our decision tree properties defined we can begin the modeling process. From the decision tree model, we looked at our first split to see which variable is the most significant in our model. We found that *Standard_Orders_Used* was the most significant variable and will be the splitting point for interpreting our model.

Right Side Split Decision Tree Analysis

Our first split begins with *Standard_Orders_Used* which is a binary Yes or *No* question. Those patients who have used a standard order would be place on the right side of the tree and would have a 15.83% probability of being readmitted on that variable alone based on our model. The right side of the tree also contains most of our records with 81,409 of the total 101,470 from our training set being on this side. Next, we find that *Procedure_Long_Desc* as our second most significant variable for the right side of the tree. Patients who had Biopsy of Bone Marrow as their procedure were 81.55% likely to readmitted back into the hospital. This leaf though contains only 710 of the patients in our dataset.

Next, we find that most of the patient in our dataset go through Diagnostic Ultrasound with 80,699 of our patients with this *Procedure Long Descripition.* Next, we find that Doctors with the ID 208572, 319034, and 2696744 have a significantly lower readmittance rate of 3.11% compared to average of the whole dataset which is 19.42%. This give us significant insight on how we can tackle the readmittance problem for the hospital system. By looking into certain doctors, we might be able to recommend the hospital system what doctors to investigate to see if there is possible malpractice occurring or possible health process that need reform.

Lastly we will go down another layer and see what nurses might be aiding in the increase readmittance rate. First we will go down the node with the doctor ID of 287,656, 319,038, and 235,415 then the node with the Diagnoistic long descripition of Acute on Chronic Diast. From there we can see what nurses could be an area of concern for patients being readmitted. The discharge nurses with the ID 145604, 281540, and 372164 had their patients readmitted 31.87% of the time.

Left Side Split Decision Tree Analysis

Just like the right side split, the first main split is based on if the standard order was used with the patient. For this side patients didn't have a standard order used and they had a 34.01% chance of being readmitted based on that variable alone

The next split is based on which doctors the patient had. If the patient had doctors with the ID 319038, 235415, or 312991 they had 44.23% chance of being readmitted compared to the doctors with the ID of 287656, 306822, and 292909 whose patients had a readmittance chance of 16.21%

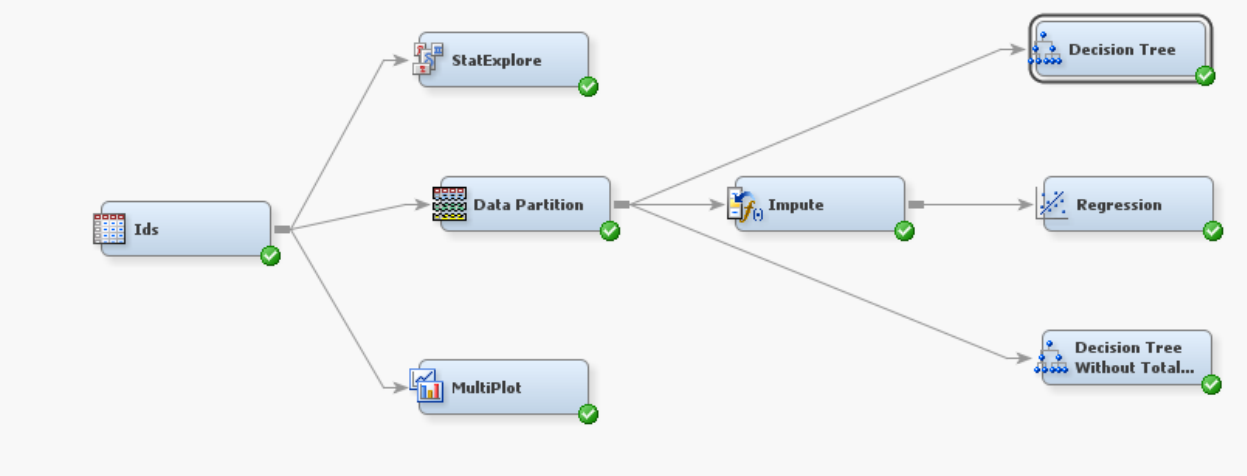
If we continue to drill down on those doctors with the 44.23% of having their patients readmitted, we gain more additional insight. Patients who had Arterial Catherization had huge increase in being

readmitted with a percentage of 95.15%. We also find that discharge nurses with ID 370015, 15006, and 270011 increase the chance of a patient being readmitted by 100%.

Lastly our model had a misclassification rate of 15.62% for our training data set and 15.98% for the validation data set. The tightness of the difference between our two data set shows that our model was trained extremely well and is ready a test dataset.

Appendix A: Screenshots

EM Diagram with Tree Node

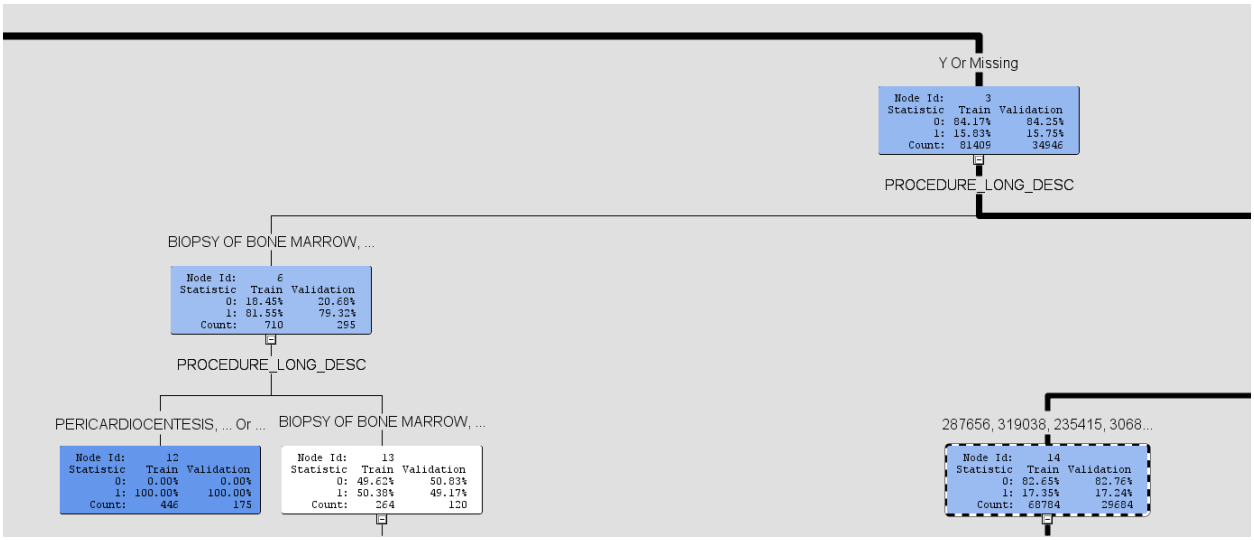


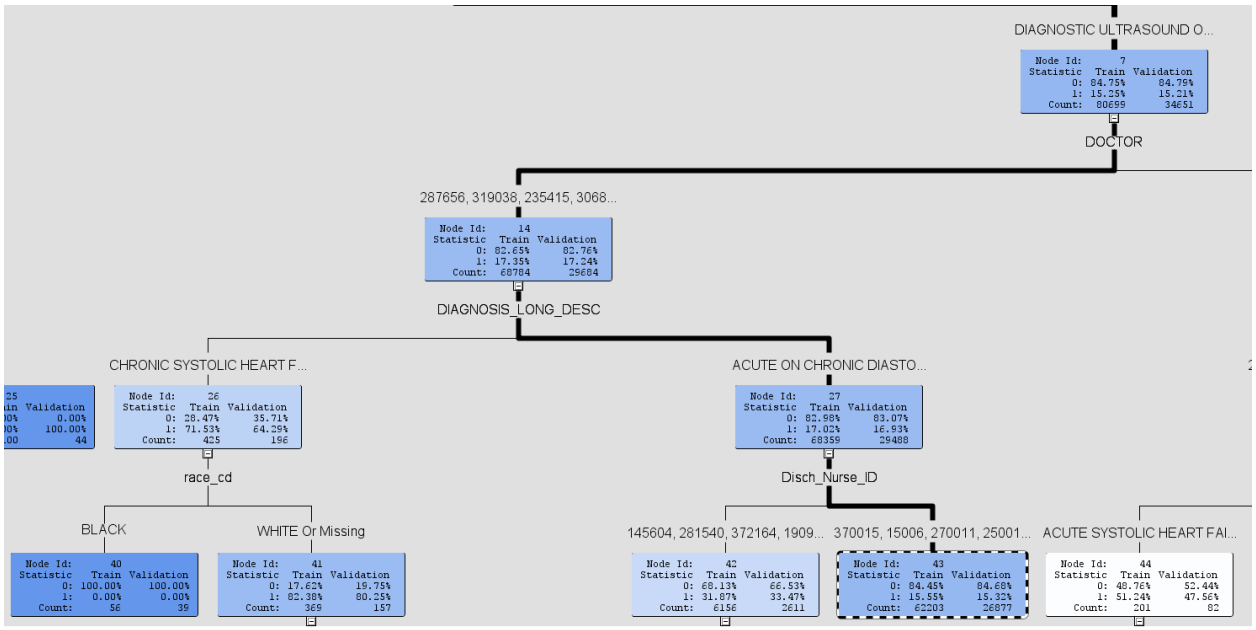
Decision Tree Properties

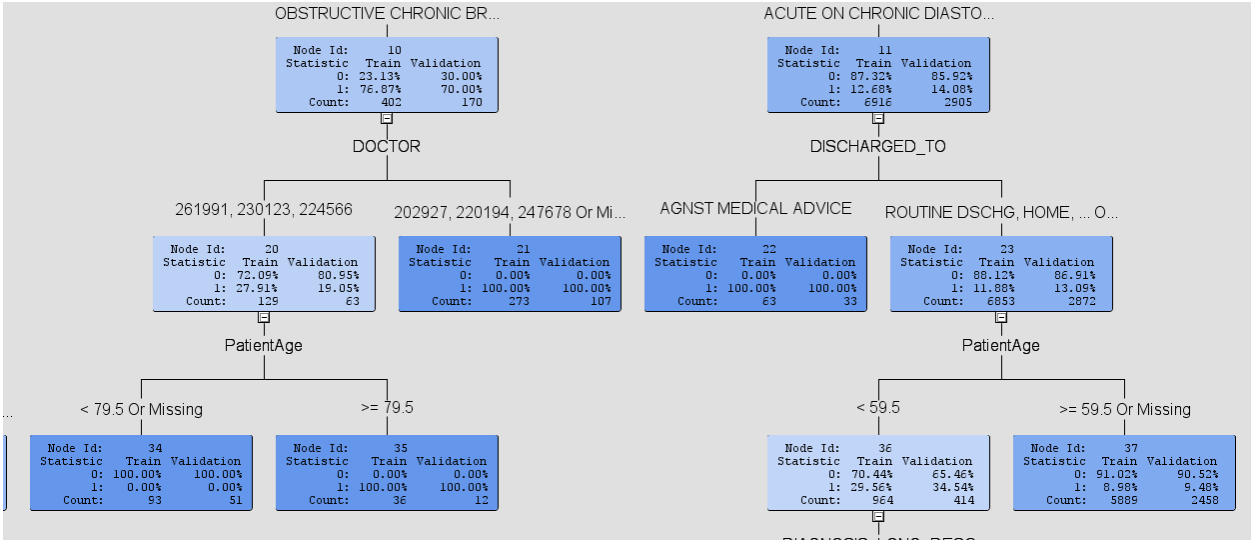
General	
Node ID	Tree2
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Interactive	
Import Tree Model	No
Tree Model Data Set	
Use Frozen Tree	No
Use Multiple Targets	No
Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categorical Size	5
Node	
Leaf Size	30
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000

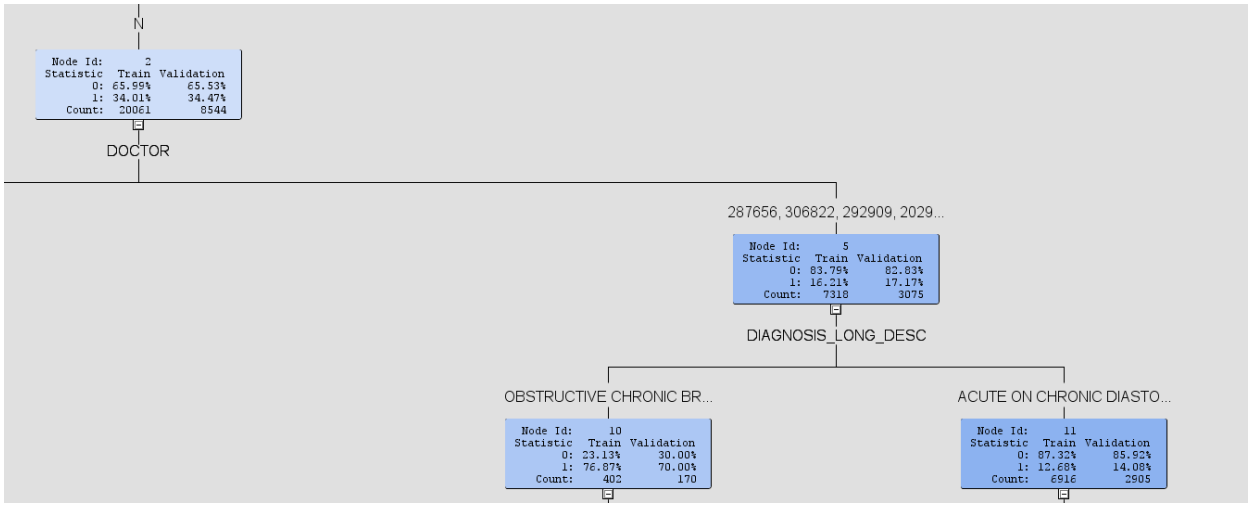
Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Decision
Assessment Fraction	0.25
Cross Validation	
Perform Cross Validation	No
Number of Subsets	10
Number of Repeats	1
Seed	12345
Observation Based Importa	
Observation Based Importa	No
Number Single Var Importar	5
P-Value Adjustment	
Bonferroni Adjustment	Yes
Time of Bonferroni Adjustm	Before
Inputs	No
Number of Inputs	1
Depth Adjustment	Yes
Output Variables	
Leaf Variable	Yes
Interactive Sample	
Create Sample	Default
Sample Method	Random
Sample Size	10000
Sample Seed	12345
Performance	Disk
Score	
Variable Selection	Yes
Leaf Role	Segment
Report	
Precision	4
Tree Precision	4
Class Target Node Color	Percent Correctly Classified
Interval Target Node Color	Average
Node Text	

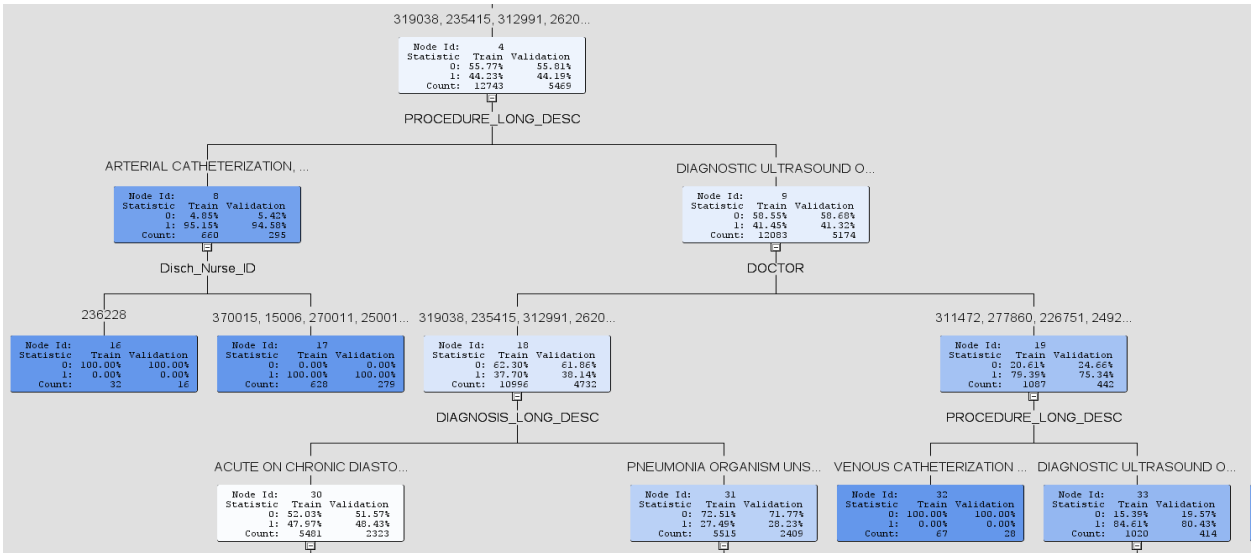
Decision Tree

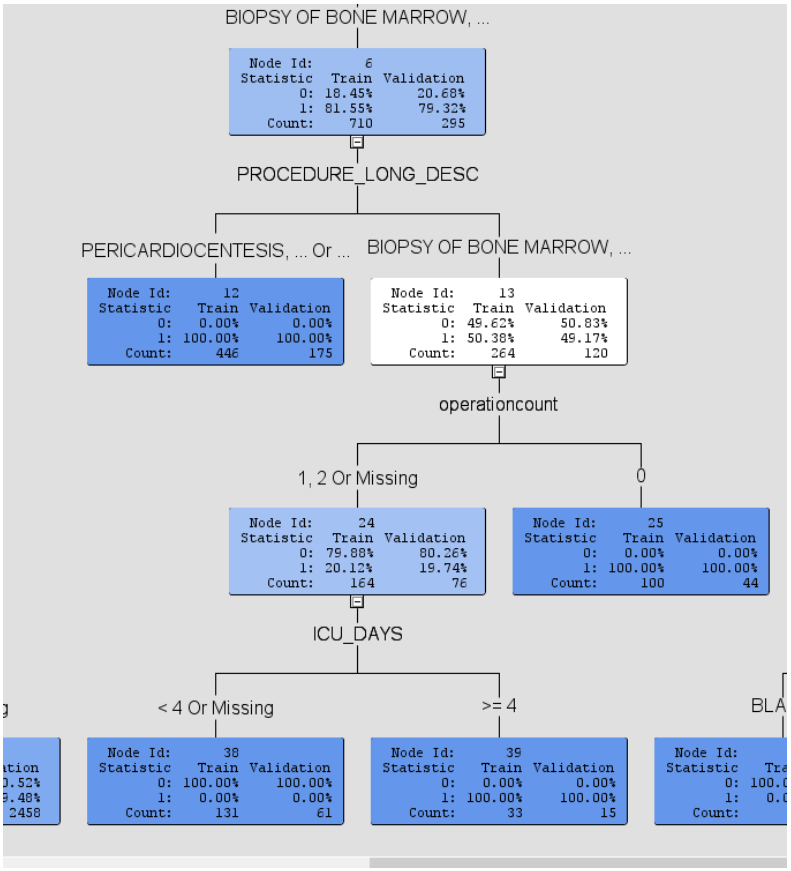




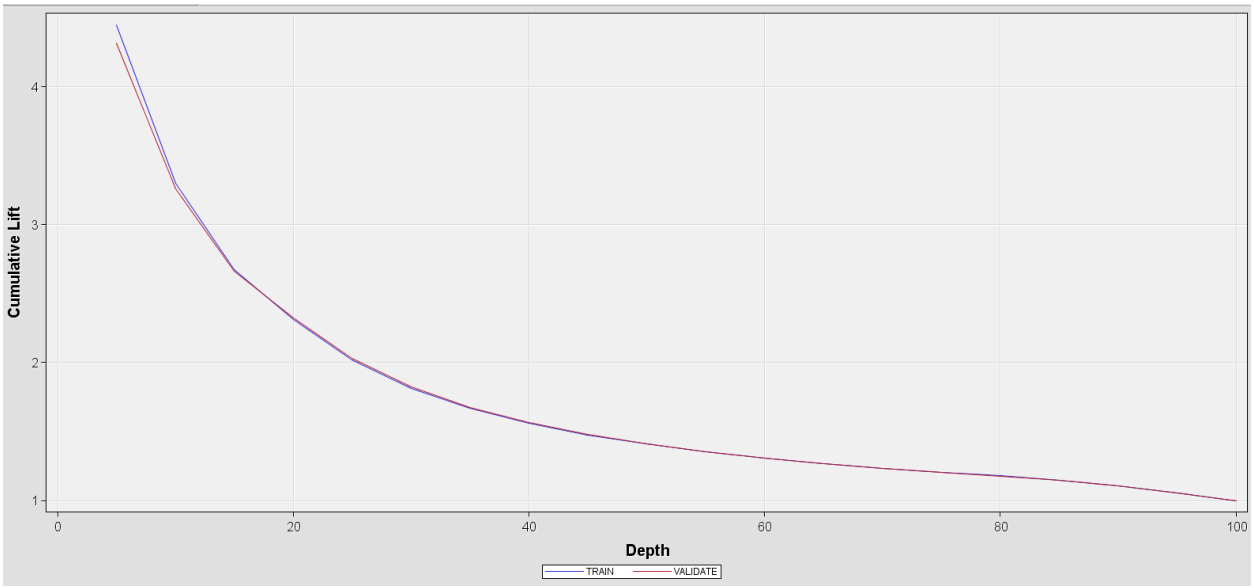

OBSTRUCTIVE CHRONIC BR...
Node Id: 10
Statistic Train Validation
0: 23.13% 30.00%
1: 76.87% 70.00%
Count: 402 170
DOCTOR
261991, 230123, 224566
Node Id: 20
Statistic Train Validation
0: 72.09% 80.95%
1: 27.91% 19.05%
Count: 129 63
202927, 220194, 247678 Or Mi...
Node Id: 21
Statistic Train Validation
0: 0.00% 0.00%
1: 100.00% 100.00%
Count: 273 107
PatientAge
< 79.5 Or Missing
Node Id: 34
Statistic Train Validation
0: 100.00% 100.00%
1: 0.00% 0.00%
Count: 93 51
>= 79.5
Node Id: 35
Statistic Train Validation
0: 0.00% 0.00%
1: 100.00% 100.00%
Count: 36 12
ACUTE ON CHRONIC DIASTO...
Node Id: 11
Statistic Train Validation
0: 87.32% 85.92%
1: 12.68% 14.08%
Count: 6916 2905
DISCHARGED_TO
AGNST MEDICAL ADVICE
Node Id: 22
Statistic Train Validation
0: 0.00% 0.00%
1: 100.00% 100.00%
Count: 63 33
ROUTINE DSCHG, HOME, ... O...
Node Id: 23
Statistic Train Validation
0: 88.12% 86.91%
1: 11.88% 13.09%
Count: 6853 2872
PatientAge
< 59.5
Node Id: 36
Statistic Train Validation
0: 70.44% 65.46%
1: 29.56% 34.54%
Count: 964 414
>= 59.5 Or Missing
Node Id: 37
Statistic Train Validation
0: 91.02% 90.52%
1: 8.98% 9.48%
Count: 5889 2458

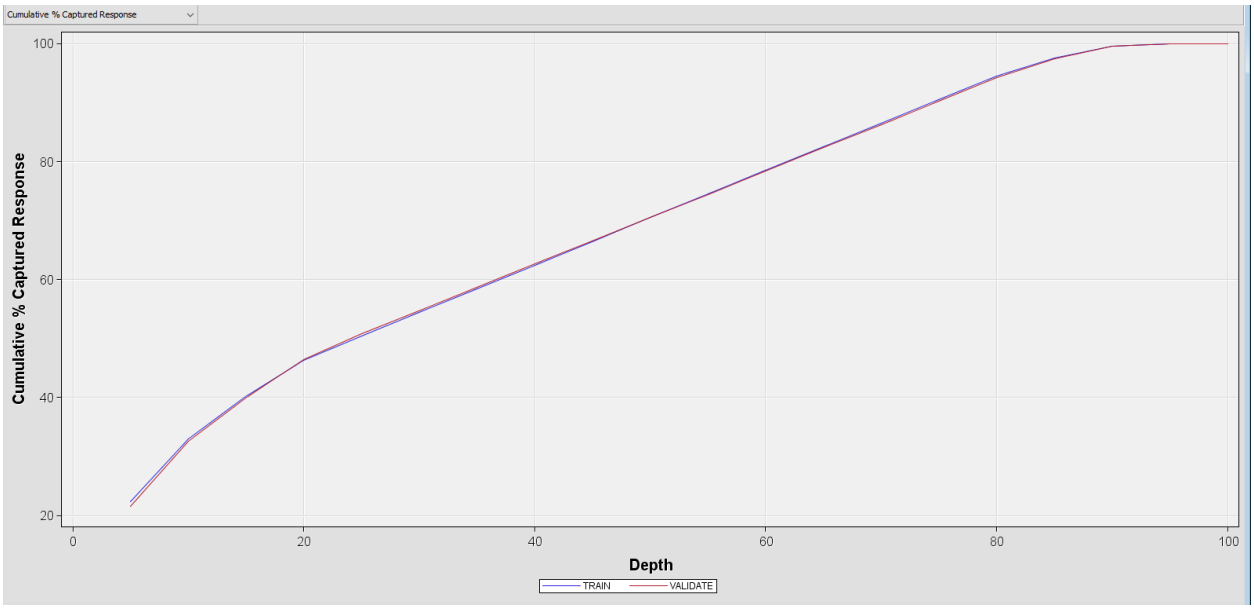

N
Node Id: 2
Statistic Train Validation
0: 65.99% 65.53%
1: 34.01% 34.47%
Count: 20061 8544
DOCTOR
287656, 306822, 292909, 2029...
Node Id: 5
Statistic Train Validation
0: 83.79% 82.83%
1: 16.21% 17.17%
Count: 7318 3075
DIAGNOSIS_LONG_DESC
OBSTRUCTIVE CHRONIC BR...
Node Id: 10
Statistic Train Validation
0: 23.13% 30.00%
1: 76.87% 70.00%
Count: 402 170
ACUTE ON CHRONIC DIASTO...
Node Id: 11
Statistic Train Validation
0: 87.32% 85.92%
1: 12.68% 14.08%
Count: 6916 2905


319038, 235415, 312991, 2620...
Node Id: 4
Statistic Train Validation
0: 55.77% 55.81%
1: 44.23% 44.19%
Count: 12743 5469
PROCEDURE_LONG_DESC
ARTERIAL CATHETERIZATION, ...
Node Id: 8
Statistic Train Validation
0: 4.85% 5.42%
1: 95.15% 94.58%
Count: 660 295
Disch_Nurse_ID
236228
Node Id: 16
Statistic Train Validation
0: 100.00% 100.00%
1: 0.00% 0.00%
Count: 32 16
370015, 15006, 270011, 25001...
Node Id: 17
Statistic Train Validation
0: 0.00% 0.00%
1: 100.00% 100.00%
Count: 628 279
DIAGNOSTIC ULTRASOUND O...
Node Id: 9
Statistic Train Validation
0: 58.55% 58.68%
1: 41.45% 41.32%
Count: 12083 5174
DOCTOR
319038, 235415, 312991, 2620...
Node Id: 18
Statistic Train Validation
0: 62.30% 61.86%
1: 37.70% 38.14%
Count: 10996 4732
DIAGNOSIS_LONG_DESC
ACUTE ON CHRONIC DIASTO...
Node Id: 30
Statistic Train Validation
0: 52.03% 51.57%
1: 47.97% 48.43%
Count: 5481 2323
PNEUMONIA ORGANISM UNS...
Node Id: 31
Statistic Train Validation
0: 72.51% 71.77%
1: 27.49% 28.23%
Count: 5515 2409
311472, 277860, 226751, 2492...
Node Id: 19
Statistic Train Validation
0: 20.61% 24.66%
1: 79.39% 75.34%
Count: 1087 442
PROCEDURE_LONG_DESC
VENOUS CATHETERIZATION ...
Node Id: 32
Statistic Train Validation
0: 100.00% 100.00%
1: 0.00% 0.00%
Count: 67 28
DIAGNOSTIC ULTRASOUND O...
Node Id: 33
Statistic Train Validation
0: 15.39% 19.57%
1: 84.61% 80.43%
Count: 1020 414



Cumulative Lift & Cumulative % capture Response Charts





Accuracy Measure

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
readmit_number		_NOBS_	Sum of Frequencies	101470	43490
readmit_number		_MISC_	Misclassification Rate	0.157298	0.159899
readmit_number		_MAX_	Maximum Absolute Error	0.945736	0.945736
readmit_number		_SSE_	Sum of Squared Errors	25262.42	10937.3
readmit_number		_ASE_	Average Squared Error	0.124482	0.125745
readmit_number		_RASE_	Root Average Squared Error	0.35282	0.354605
readmit_number		_DIV_	Divisor for ASE	202940	86980
readmit_number		_DFT_	Total Degrees of Freedom	101470	.