Executive Summary

Introduction: Searching for the best model

What type of model has the most predictive power in predicting who will be readmitted to the hospital?

For our model comparison on our four models, two of which were decision trees and the other two were logistic regression we found the original decision tree model to be the best model for predicting patient re-admittance. It had the lowest misclassification rate from all the other models with it an error rate of 11.27% compared to the second best model which was the Logistic Regression with the four interaction which had a misclassification rate of 14.54%.

Is there specific business insight in the non-winning model type that would lead you to use that model for scoring? What is that insight?

The decision tree model while not our best model based on predictability does provide a much more business-centric model. It follows a very business process focus way of thinking where the splits were done based on how a patient would enter a hospital. The value from this model is the ease of interpretability for people within the business and also the fact that it focuses more on the variables that are often used within the hospital.

The logistic regression model with the interaction provided us additional insight into how our specific explanatory variables interact with each other. Our hypothesis on these explanatory variables having interactions with each was Operation Count and Patient Age, Days in ICU and Length of Stay, Diagnosis Group and Length of Stay, and Procedure Long Description and Patient Age. From running our model, we found that our hypothesis was correct and there was an unseen relationship that needs more investigation which can provide additional business insights. These interactions improved upon our original logistic regression model which can be seen in Appendix B.

On choosing which model to use it depends on what is valued more. If the business wants higher predictive power the best model would be the decision tree which has an error rate of 3% better than the second-best model. Now if a business-centric and understandable model is wanted the decision tree - Interactive would be the best model as it follows a more business process model, but it has a substantially higher error rate.

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 systems who have a much lower admittance 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.

With the new logistic regression and decision tree model, the next step is to compare the four models. Using the *Model Comparison Node* we set the properties to score on misclassification rate as our model is binary. The four models being compared are our original logistic and decision tree model and the new interactive decision tree and logistic regression with an interaction. The new decision tree is based on a more business-centric approach with Diagnosis Long Description being the main split for the model. The next two splits before the model were trained were Procedure Long Description and Doctor. The purpose of this tree is to follow the business process of a hospital. Lastly, the new logistic regression model is also based on a more business-centric approach. Our team came up with multiple hypotheses for our explanatory variables and the relationship they may have with each other.

Key Findings

From our decision interactive tree, we found that we can still have of a somewhat of an effective model when we focus it on a more business-centric approach. The interactive tree gave us additional insight into how some our explanatory interact with each other. What we found is the following explanatory variables had a relationship amongst each other which improved our model:

- Operation Count and Patient Age
- Days in ICU and Length of Stay
- Diagnosis Group and Length of Stay
- Patient Age and Procedure Long Description

When it came to our model comparison we were able to determine the best model when it came to predicting if a patient would be readmitted. The best model was the original decision tree, followed by the logistic regression with interaction, the original logistic regression, and the interactive decision tree respectively. This is based on using the misclassification rate as our main metric, where the original decision tree had an error rate of about 11.2% compared to the other three models which were hovering between 14% - 15%. When using the cumulative lift chart, while our winning model is still the best it starts to see its value start to diminish at 20% decile. After the 20% most likely patients are identified the winning model is no better in terms of predictive value than the other three models.

Recommendations

For the model that the hospital should use in production, our data science team recommends the original decision tree. It has the best predictive power of the four models and it has the lowest misclassification rate. Using this model will help the hospital reduce its patient admittance rate the most effectively whiling for better resource allocation.

If the hospital prefers a more business-centric model, our data science team recommends the interactive decision tree. While it was the least effective model of the four it still carries predictive value much stronger than random guessing. This model follows a more business-oriented process which can easier to interpret and also use in the business setting.

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

Introduction

For this case, the main focus is on comparing different types of models and deciding the best model from the validation dataset. From the previous cases, we created a logistic regression and decision tree model to predict the probability of a patient being readmitted back into the hospital based on the variables chosen from our data set. Our data science team has decided to create two additional models to see if we can create a better model or a model that fits the business problem. A additional Logistic regression and decision tree were created that used the same variable but with different properties used. With the additional models created the *Model Comparison Node* was used to compare all four model where our data science team was able to conduct additional analysis on choosing the best model.

Additional Model Creation

Logistic Regression - Interaction

Just like the first Logistic many of the same properties were kept such as:

- **Selection Model:** Stepwise
- **Selection Criterion:** Validation Misclassification

For the new logistic regression, interaction was added to see if our team could be lower on the insight gained from the *Graph Explore Node.* Interactions were used to test our team's hypothesis for if there were a relationship between the explanatory variables in our model. To test our hypothesis the *terms editor* was used which allows for two explanatory variables to interact with each other in the model. The main purpose of this property is to make variables that are not significant in the previous model and make them significant in the newer model. The Interactions that were used in the new logistic regression model were:

- **DAYS_ICU * LENGTH_OF_STAY:** Our team's hypothesis on this interaction is that the more time you spend in the ICU the patient should be expected to stay there longer. There is a slight correlation that can be seen from the *Graph Explore Node*. The graph can be found in Appendix B.
- **Operationcount * PatientAge**: Our team's hypothesis on this interaction is that as the older the patient is, the more operations they will need as their bodies don't recover as fast as someone who is much younger. There is a slight correlation that can be seen from the *Graph Explore Node.* The graph can be found in Appendix B.
- **LENGTH_OF_STAY * Diagnosis_Group:**
- **PatientAge * PROCEDURE_LONG_DESC:**

Decision Tree - Interactive

Just like the logistic regression model most of the properties used in the interactive decision tree derived from the original model. Some of those properties include:

- **Minimum Leaf Size:** The maximum leaf size was change from 30 to 100 to be more strict on the leaf nodes
- **Maximum Depth:** The maximum depth was left again at the default of 6 to ensure a simple interpretable model
- **Maximum Branch:** The maximum branch was left again at the default of 2 to ensure a simple interpretable model
- **Significance Level:** the significance level was set at 0.2 as we are still working with a decently large dataset

For this model, we use the property called *Interactive* on the *Decision Tree Node.* This property allows us to choose the splits for each depth while also training the model on our chosen leaves. This is helpful for fitting the model to the specific business needs and requirements that statistics itself can't do.

On the interactive tree, the first step our team did was prune the tree of the previously trained model. This allows us to choose the first split for the model. Our team went through different variables such as ***days_ICU, procedure***, and ***length_of_stay*** to see the value that they added statistically and to the business. From the various variables, the variable that was chosen was ***diagnosis_long_desc*** as our best first split. From this split, we again manually split the leaf nodes again using ***procedure_long_desc*** as our second split and ***doctor*** as our third split. For the reasoning behind the variables chosen for the splits in the decision tree, we will begin with the first variable. When patients come into the hospital one of the first things that are done to them is they are diagnosed with a condition and then a procedure is conducted to help the patient. If our model finds that there is a certain diagnosis that are more prone to readmittance the hospital team can move resources more efficiently and quickly from the very moment the patient arrives. Lastly, for the third split, the variable doctor was chosen as a patient is assigned a doctor for a certain procedure which helps gain insight of doctors who cause more problems.

Additional Enterprise Miner Nodes

Graph Explore Node

The Graph Explore Node was added into the Enterprise Miner data mining diagram to help us conduct additional exploratory data analysis. With the new logistic regression model using interactions additional insight had to be gather in order to create new hypothesis for how certain explanatory variables might affect each other. This node is connected to the impute node as it will allow us to work with a dataset without missing values. All the properties were left as default and from there multiple scatter plots were created to see if there see if there were any interactions between the variables of use in our models.

Model Comparison

With all the models created the last and final step to do is to compare all of them based on their predictive power. To do this the *Model Comparison Node* was used to assess the prediction from the preceding modeling nodes. All the models created were then connected to this node allowing for them to be compare. Before running the model some of the default properties were changed. For the *Selection Table* property, it was changed to validation since it's the dataset that was partition to test the effectiveness of

the model trained on the training partition dataset. The next property that was change was *Selection Statistic.* The property field was switch to Misclassification Rate since the variable of interest is binary

Analysis/ Finding

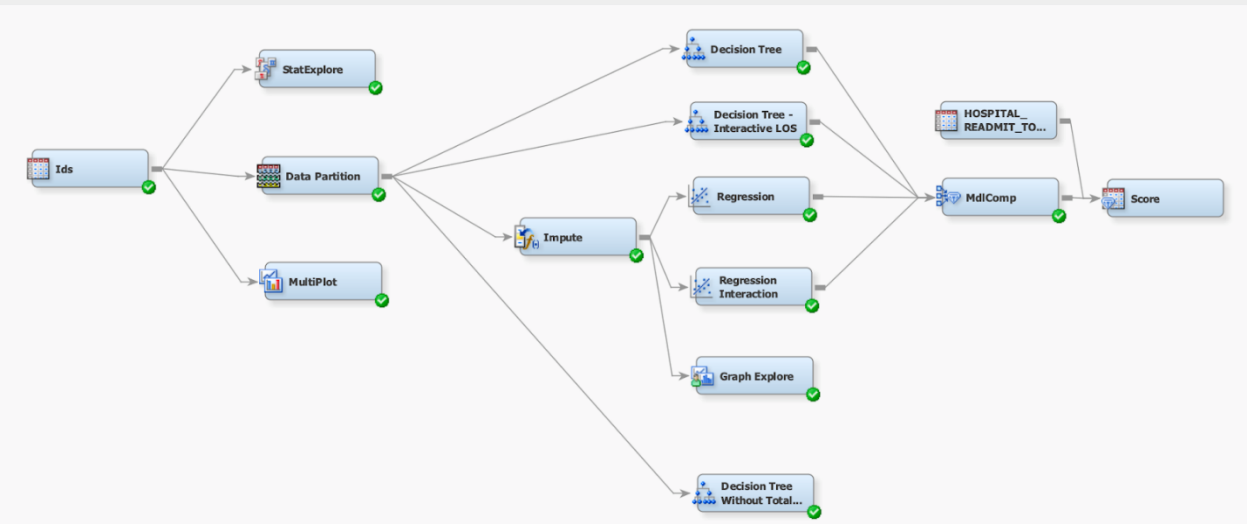
After the model comparison node was ran we were able to find the ranking of our models based on their predictive model. The ranking of the models based on the Misclassification Rates are:

- **Original Decision Tree:** 0.1127 (Winning Model)
- **Logistic Regression with Interaction**: 0.1454
- **Original Logistic Regression**: 0.1507
- **Decision Tree** - Interactive: 0.1612

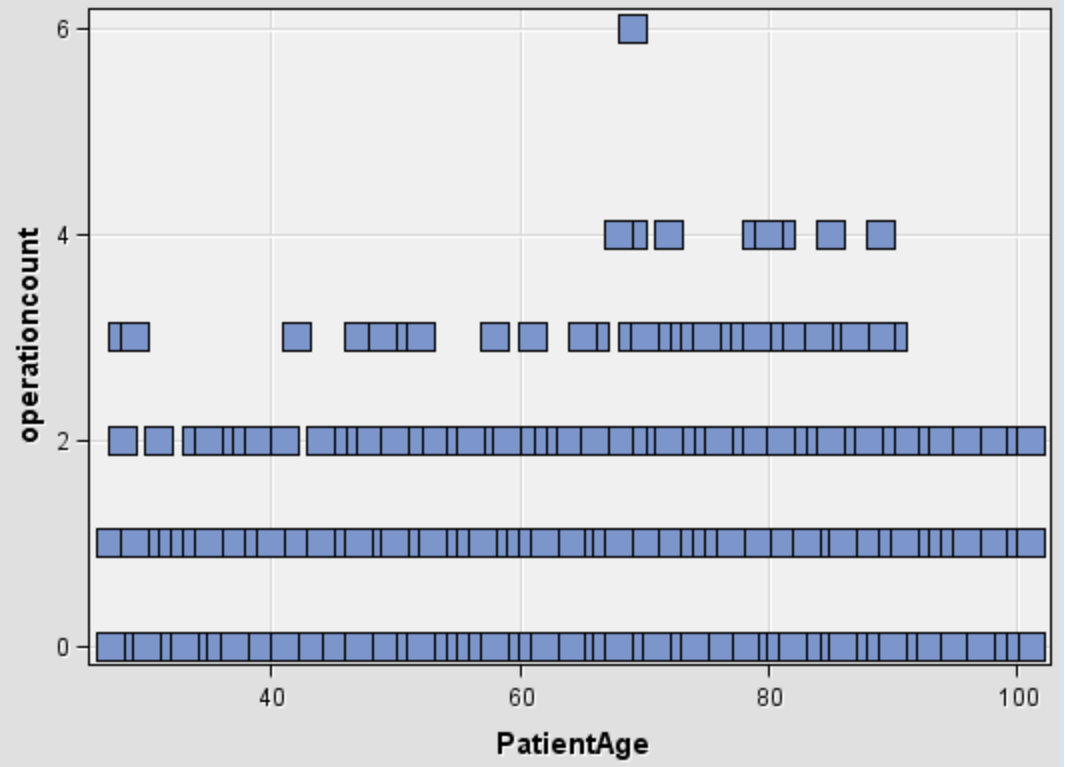
The rankings for these models also stay consistent when looking at other model comparison charts such as the ROC Chart and the Cumulative Lift Chart. While the rankings are consistent there are some interesting findings from those charts that could be found useful. After the 20% decile we start to see a huge drop off in value from our winning model. There is a decent amount of separation from our original decision tree model to the other models up until the 16% depth into our dataset. What we can conclude from this is that the decision tree model is very useful when it comes to predicting those patients with very high probability of being readmitted but starts to fall short the more patients that it predicts correctly. We also found that the logistic model with the interaction perform slightly better than the original. This shows that our hypothesis was correct and provided additional lift to the model.

Appendix B - Screenshots

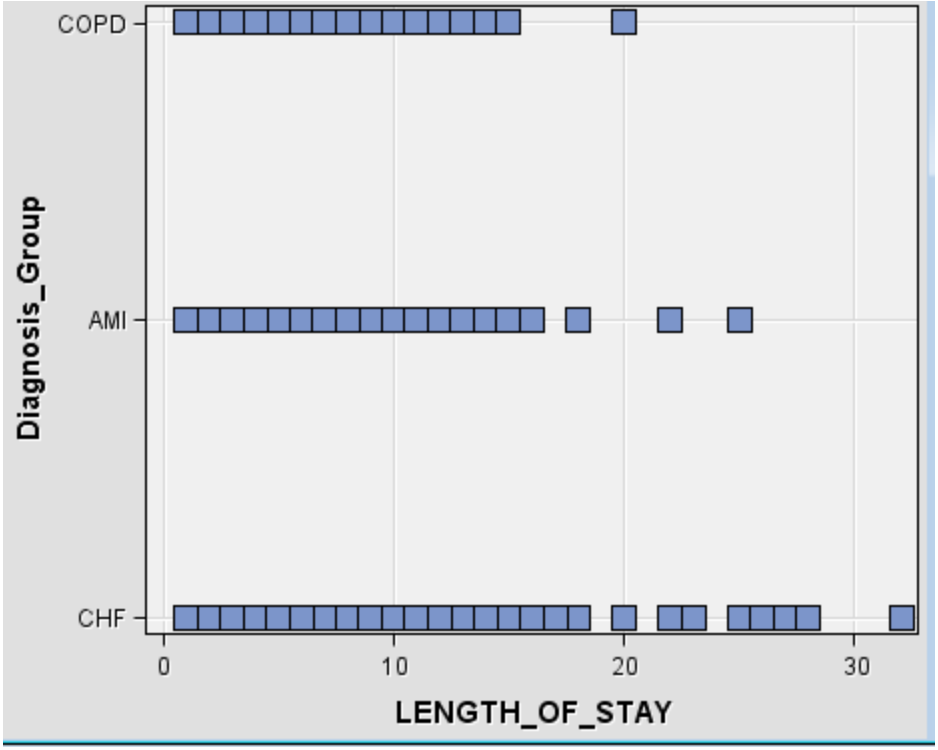
Diagram:



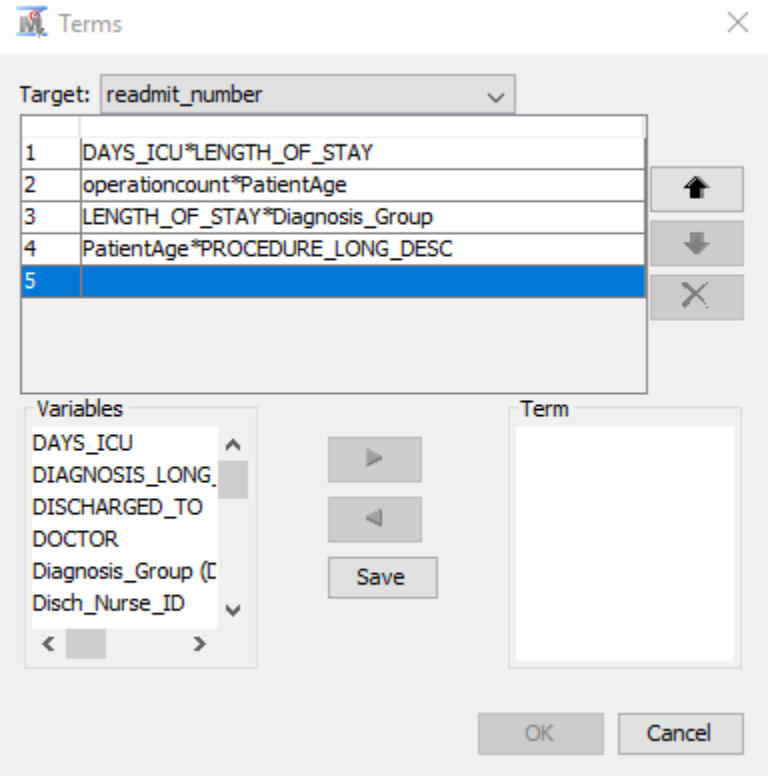
Graph Explore: PatientAge*OperationCount



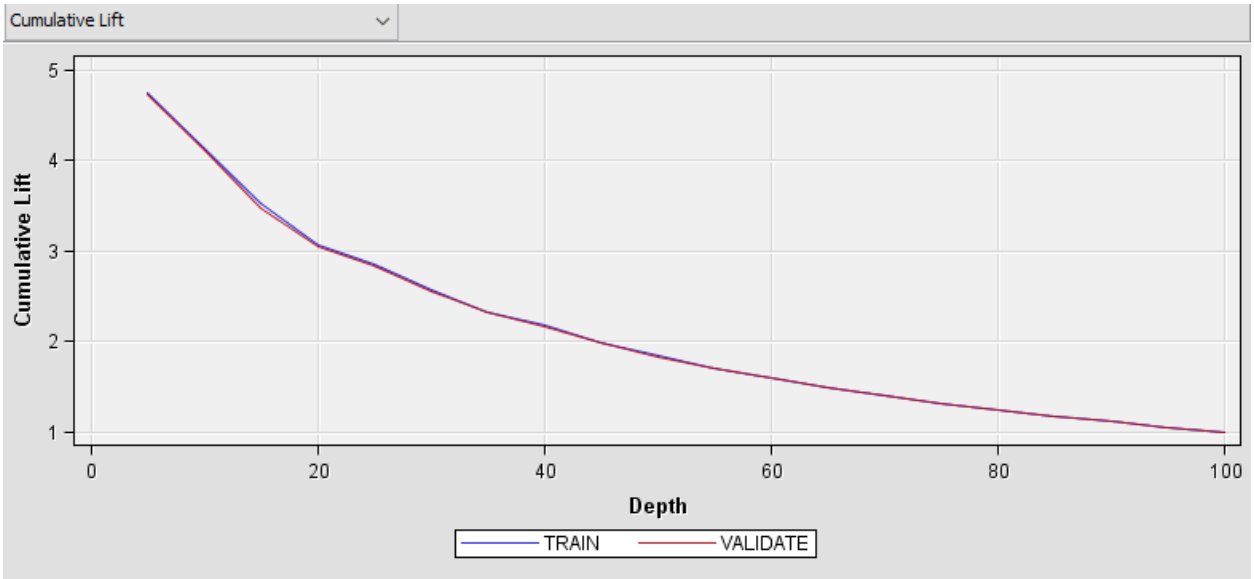
Graph Explore: Diagnosis_Group*Length_Of_Stay



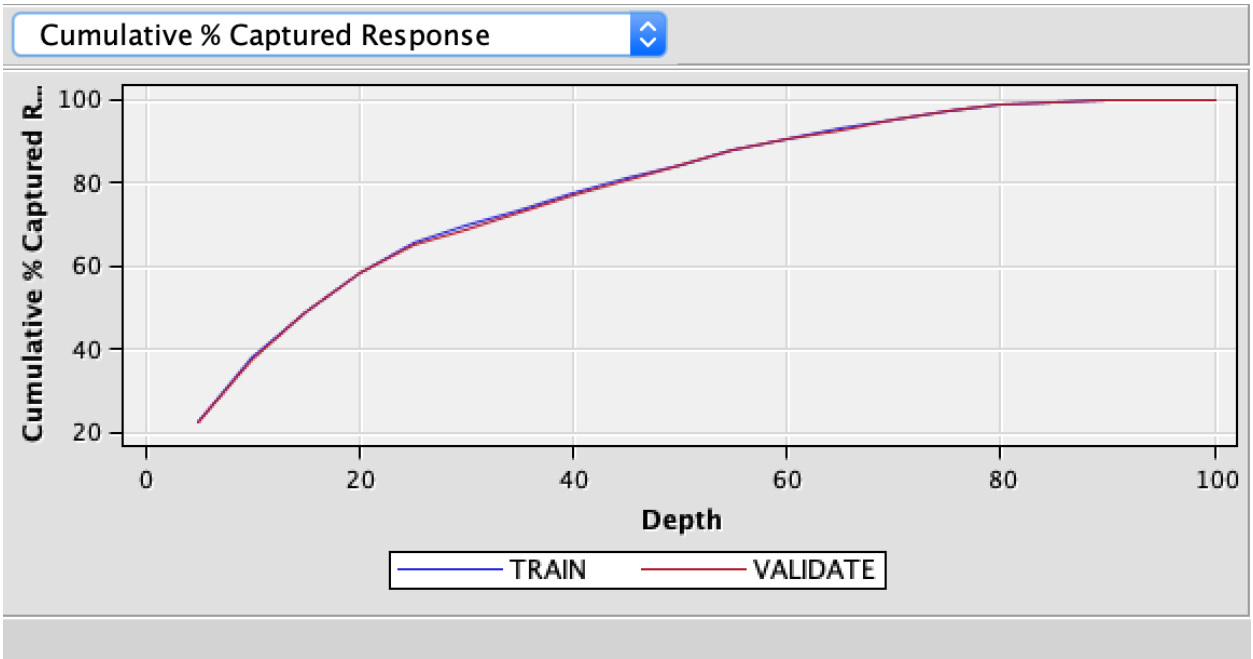
Regression Interaction: Interaction



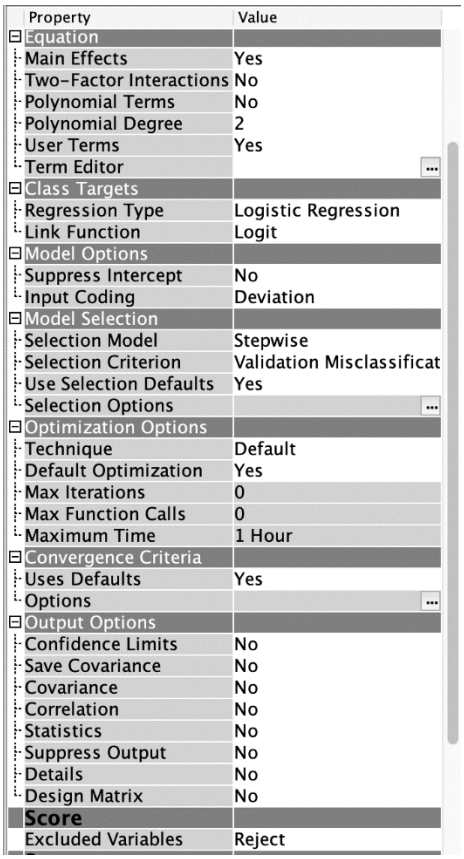
Logistic Regression Interaction: Lift Chart



Logistic Regression: Cumulative % Captured Response

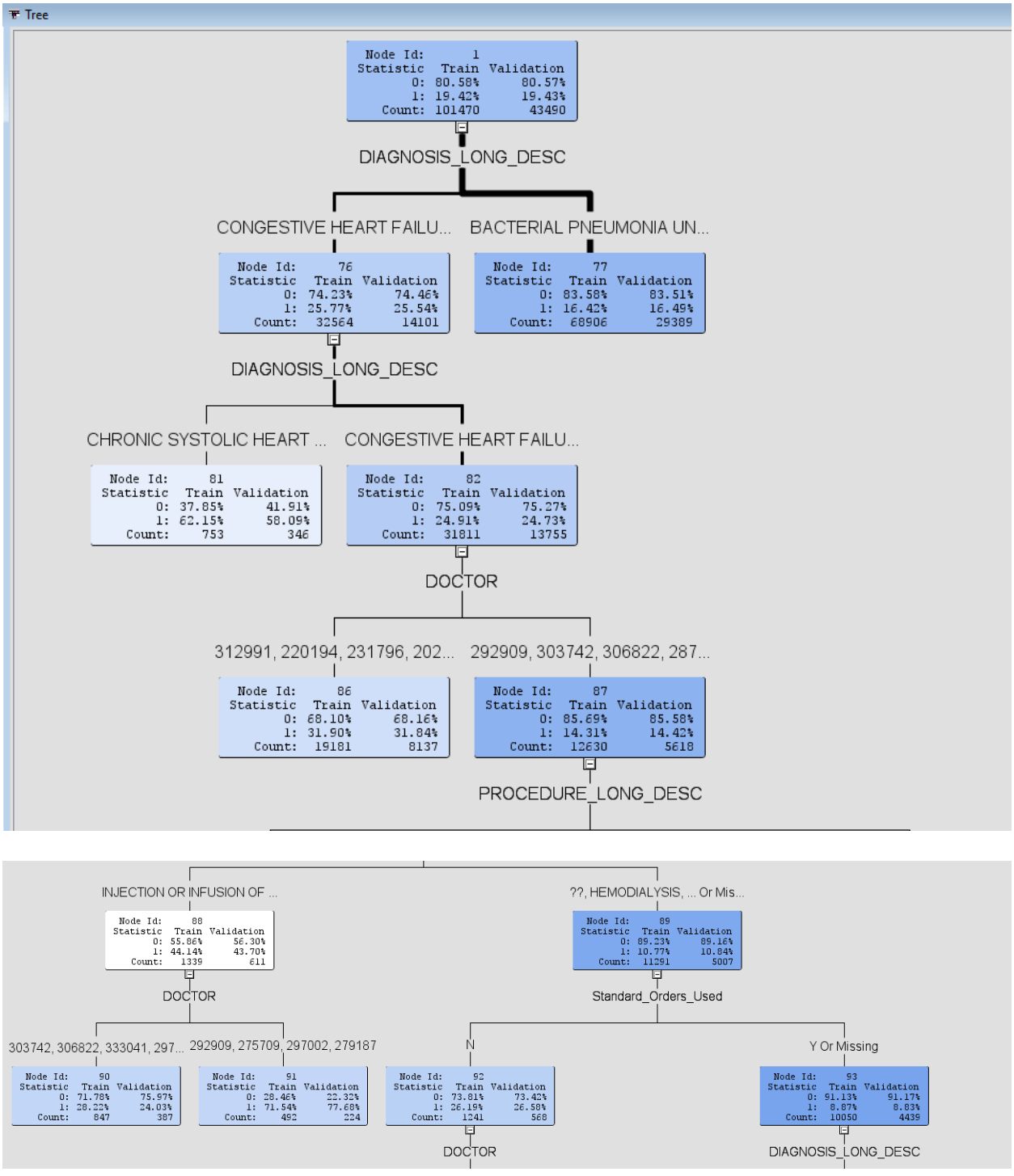


Logistic Regression: Properties

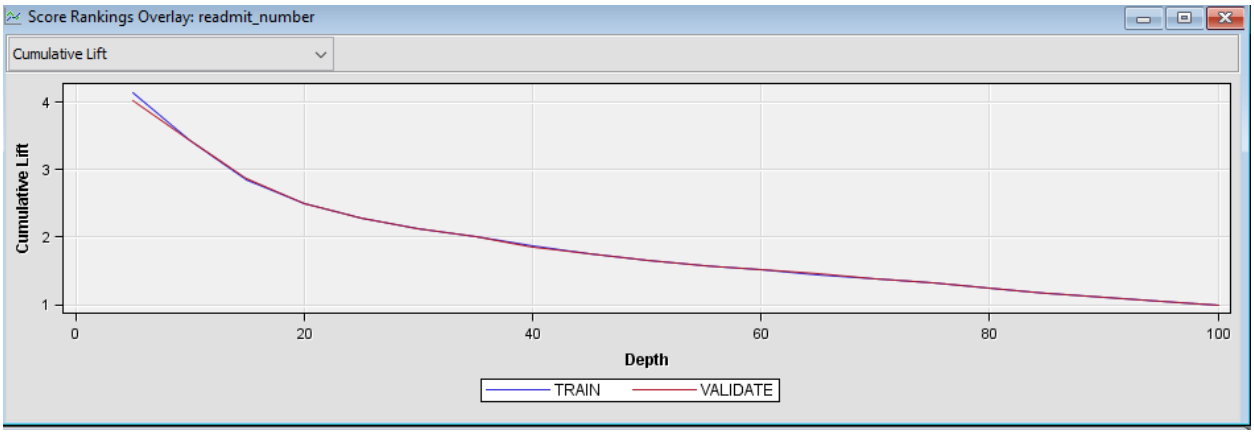


Property	Value
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	Yes
Term Editor	...
Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Stepwise
Selection Criterion	Validation Misclassificat
Use Selection Defaults	Yes
Selection Options	...
Optimization Options	
Technique	Default
Default Optimization	Yes
Max Iterations	0
Max Function Calls	0
Maximum Time	1 Hour
Convergence Criteria	
Uses Defaults	Yes
Options	...
Output Options	
Confidence Limits	No
Save Covariance	No
Covariance	No
Correlation	No
Statistics	No
Suppress Output	No
Details	No
Design Matrix	No
Score	
Excluded Variables	Reject

Decision Tree: Interactive Tree Snapshot



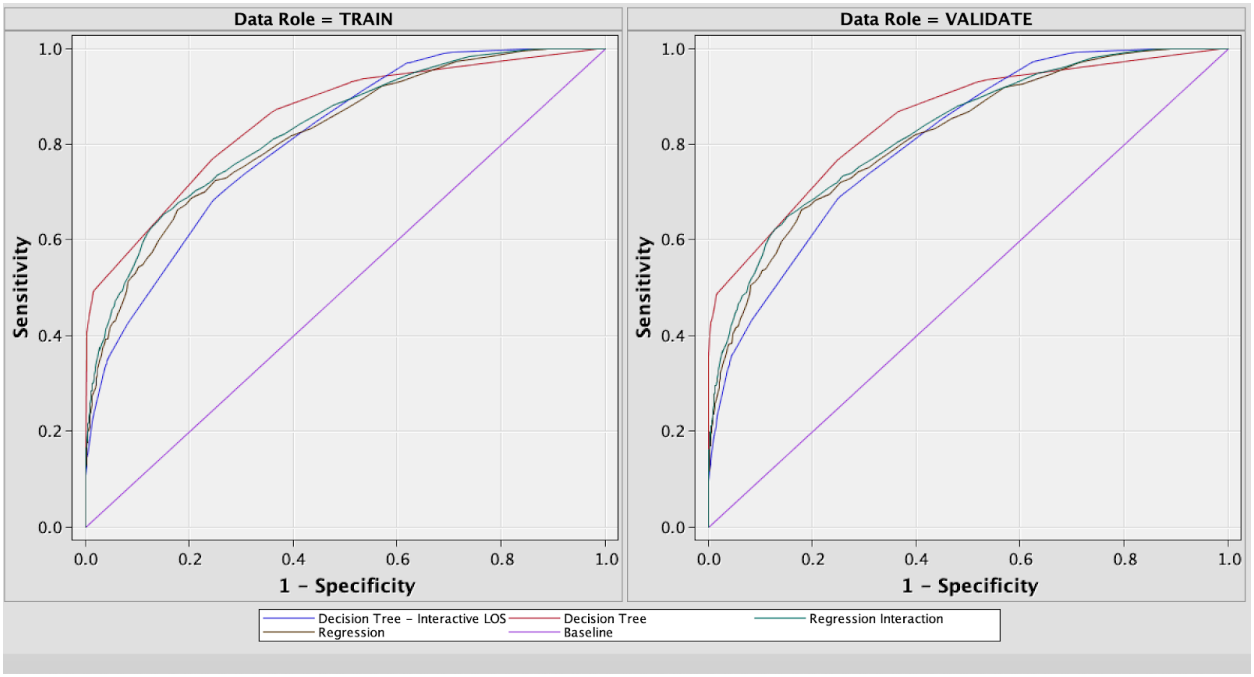
Decision Tree: Lift Chart



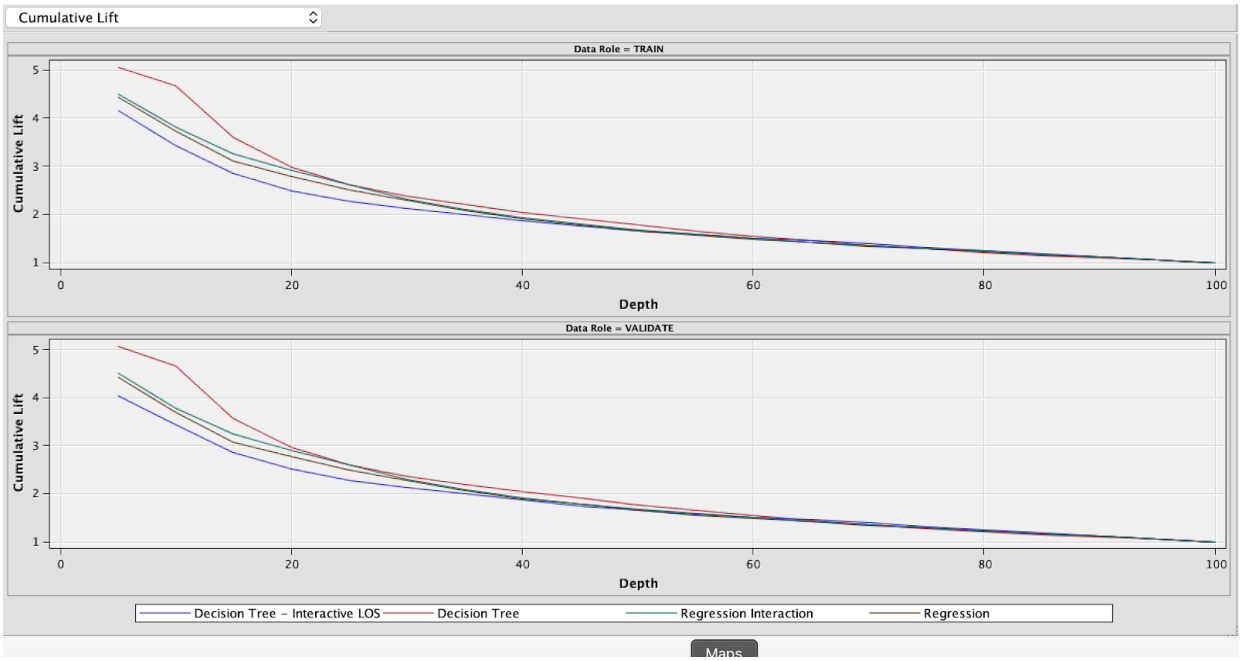
Model Comparison Node: Properties

Property	Value
General	
Node ID	MdlComp
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Assessment Reports	
Number of Bins	20
ROC Chart	Yes
Recompute	No
Model Selection	
Selection Data	Default
Selection Statistic	Misclassification Rate
HP Selection Statistic	Default
SAS Viya Selection Statis	...
Selection Table	Validation
Selection Depth	10
Score	
Selection Editor	...
Report	
Selected Model	
Target	readmit_number
Model Node	Tree
Model Description	Decision Tree
Selection Criteria	Valid: Misclassification F
Status	

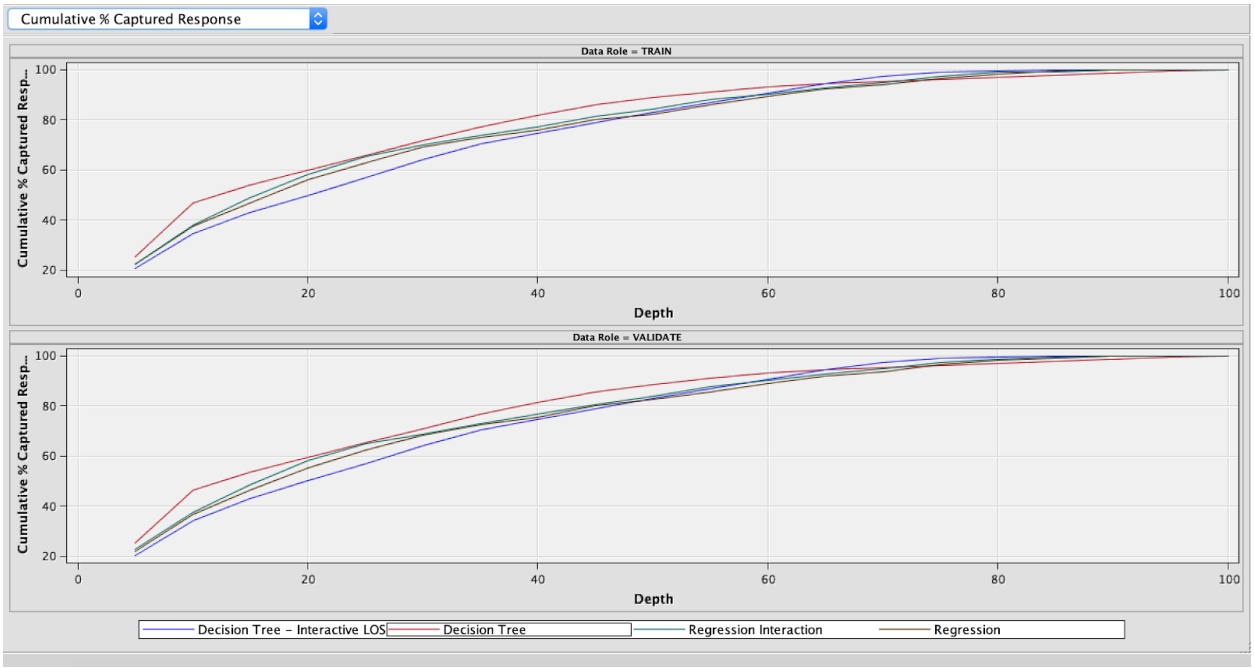
Model Comparison: ROC Chart



Model Comparison: Cumulative Lift



Model Comparison: Cumulative % Captured Response



Model Comparison: Accuracy Measure

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate
Y	Tree	Tree	Decision Tree	readmi...		0.112762
	Reg2	Reg2	Regression Interaction	readmi...		0.145482
	Reg	Reg	Regression	readmi...		0.150747
	Tree3	Tree3	Decision Tree - Interactive LOS	readmi...		0.161209