Management Report

Executive Summary: Choosing a Model and Scoring

What type of model was the EM chosen champion?

Our data science team decided to use the model comparison node within Enterprise Miner to choose the winning model for our scoring. The model that won was the original decision tree that our team created during the modeling creation process. This model has the highest predictive power compared to other 3 models that were created by a large margin. Using this model our data science team used it to score a completely new dataset that separated from the original data set that was used to train our models. An analysis was conducted to see how our model was able to predict the re-admittance probability rate on unseen patients.

Would you use a better one for the business? Which one and why?

While this model may not have the strongest predictive power the original logistic can provide interesting value when it comes to using business logic. Compared to the winning model chosen by Enterprise Miner the logistic regression model provides stronger interpretability when it comes to the business. Using this model we can use the variables that were found to be significant within the model and use it as our basis of interpretation. The hospital system can focus only on the variables that the model uses as they were found to have value in predicting whether or not they will be readmitted or not.

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 another hospital system 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.

For scoring the dataset our data science team choose to use the model that Enterprise Miner found to be the best model based on pure predictive power. The model that was chosen from the system was the original decision tree. To gain insight into how powerful this model was, the model was used to score a completely unseen dataset that the model wasn't trained on. Once the dataset was scored the newly created table was saved and then exported into a SAS library where we could conduct additional analysis on the distribution of the probabilities.

Key Findings

Once the scoring process was completed we found that the original decision tree was our best model based on predictive power. This makes sense as it had a large separation in its misclassification rate to the three other models by more than 4%. In terms of business interpretation, this model provides a new way of classifying patients who are more prone to being readmitted by choosing variables statistically. Lastly, this model had strong predictive power when it came to identifying patients who were low probability candidates when it came to being readmitted which can be seen in *Figure.6*

Recommendations: *Business Value and Potential Impact*

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

After creating our four machine learning models and comparing them using the *model comparison node* our team was tasked with using the best model to score a completely new dataset. This new dataset was created when our team was assigned to import and clean the four datasets given from the hospital. Using the unseen dataset we will see how well our model was able to do in the *scoring node* by comparing the actual readmittance number to the probability given by the model,

Input Data Node

Before we can use the *Score Node* we need a completely new dataset that hasn't been used to train and validate our models. During our data pre-processing and cleaning in SAS Visual Studio, we set aside 10% of the original dataset as its own entity to be used for scoring after the winning model was chosen. This dataset was left in the *READMIT Folder* that is currently in the cloud. To bring in this dataset into our data mining diagram we go to the data source dropdown and create a new data source This cause the data source wizard to pop-up which allows us to bring in the unseen dataset from the *READMIT* Sas library. Once we have inserted this dataset into Enterprise Miner we will drag it into our Diagram allowing it to be used for scoring. Before we can use it to score we must go into the edit variable selection box and make sure that adjust the metadata where the *readmitance_number* variable is rejected.

Score Node

Under the assess tab we drag in the *Score Node* into our diagram. This node will use the winning model from our *Model Comparison Node* and take the *HOSPITAL_READMIT_TO_SCORE_10PCT* dataset and score each record. From a scale 0-100% each record will be given a percentage of how likely they are to be readmitted back into the hospital system. Before we run the *Score Node* we need to adjust one property from this node. The "Type of Scored Data" property needs to be adjusted to *Data.* After this property is changed we will run the node and use the "Exported Data" Property of the node to compare the predicted probability to the actual binary variable.

Save Node

With our data scored the next thing we want to do is saved that newly scored dataset and analyze it. Enterprise Miner fortunately enough has a node that allows us to create a completely new SAS Table. Under the Utility tab, we drag the *Save Data Node* into the diagram. We connect this node to the *Score Node* to ensure that we saving the score dataset. Before we run this node we have to change one property to ensure that the dataset is saved in the right area. We change the "SAS Library Name" property to READMIT to ensure that our dataset is in the central location of where all our other Hospital dataset is. After the property is saved we can run the node.

Exploration: Scored-Data

With the scored dataset in the READMIT SAS library, we are now able to bring in the dataset back into the Enterprise Miner Diagram. Just like we imported the *HOSPITAL_READMIT_TO_SCORE_10PCT* dataset into the diagram we will again use the *Input Data Node* to bring in our scored dataset. Following the same steps mentioned in the Input Data Node section for the data source wizard, we will now instead choose EM_SAVE_SCORE sas table. After we inserted, the dataset into Enterprise Miner we can now drag the dataset into the diagram. With the diagram inserted we can now drag the *Graph Explore Node* into the diagram as well. We will then connect the two nodes together and run the graph explore node.

With the graph explore node run we can now go to the results and click on the plot button. With the plot, pop-up will then choose a histogram as our graph of choice. We will select the *Probability For Level 1 of Readmit_Number* as our x variable and will use percentage instead of frequency. This will result in the graph that can be found in *Appendix B: Figure 6.*

Appendix B: Screenshots

Figure 1: ReadmitData Diagram

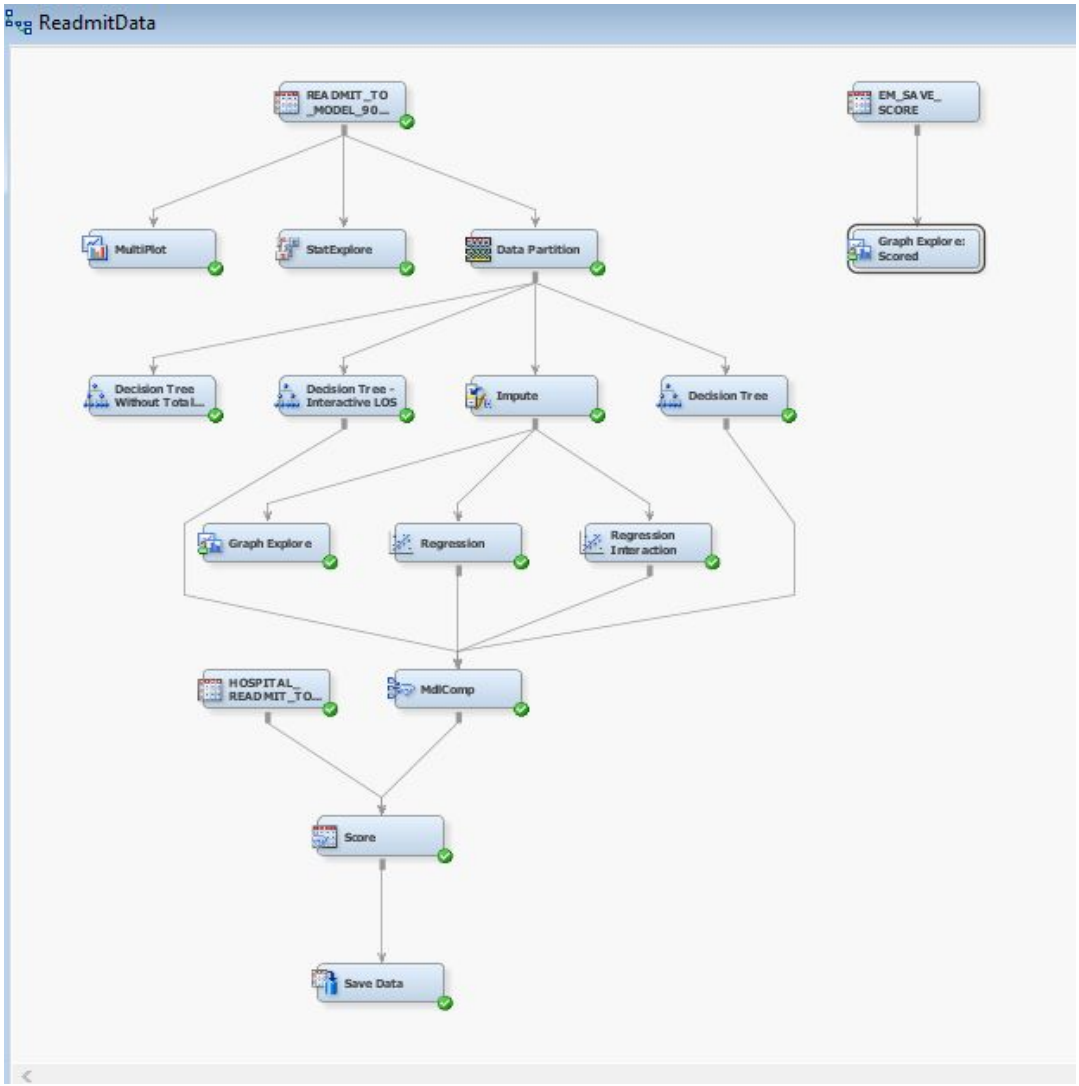
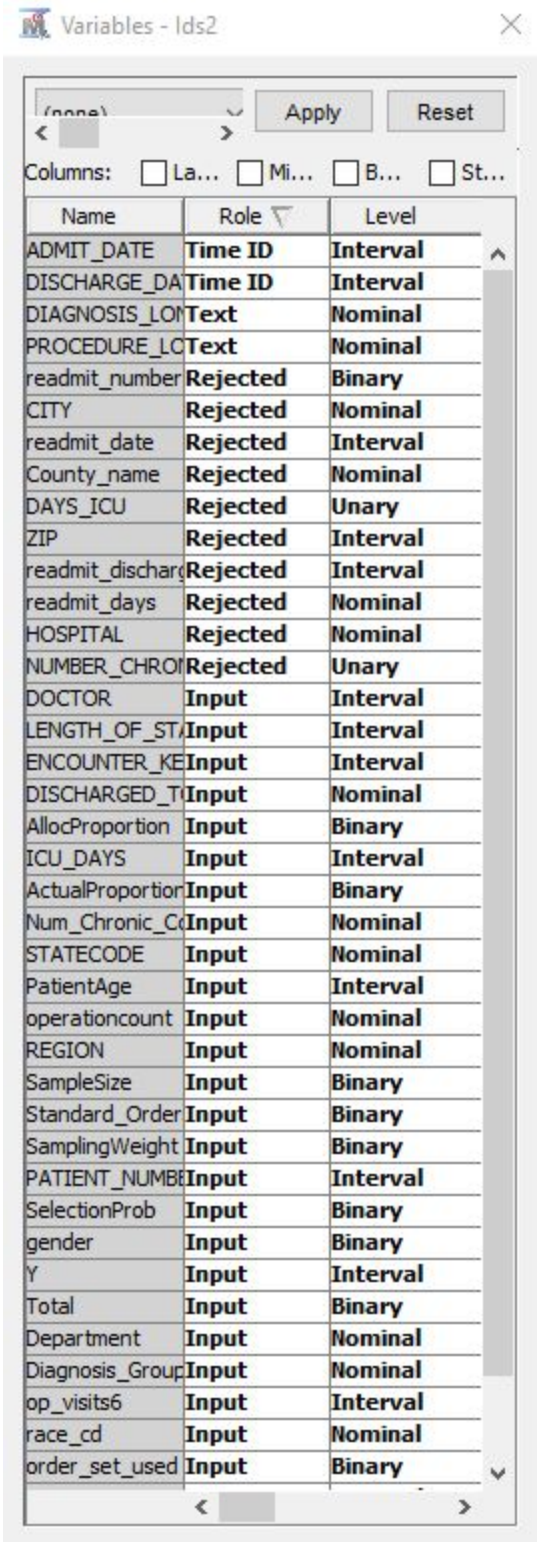
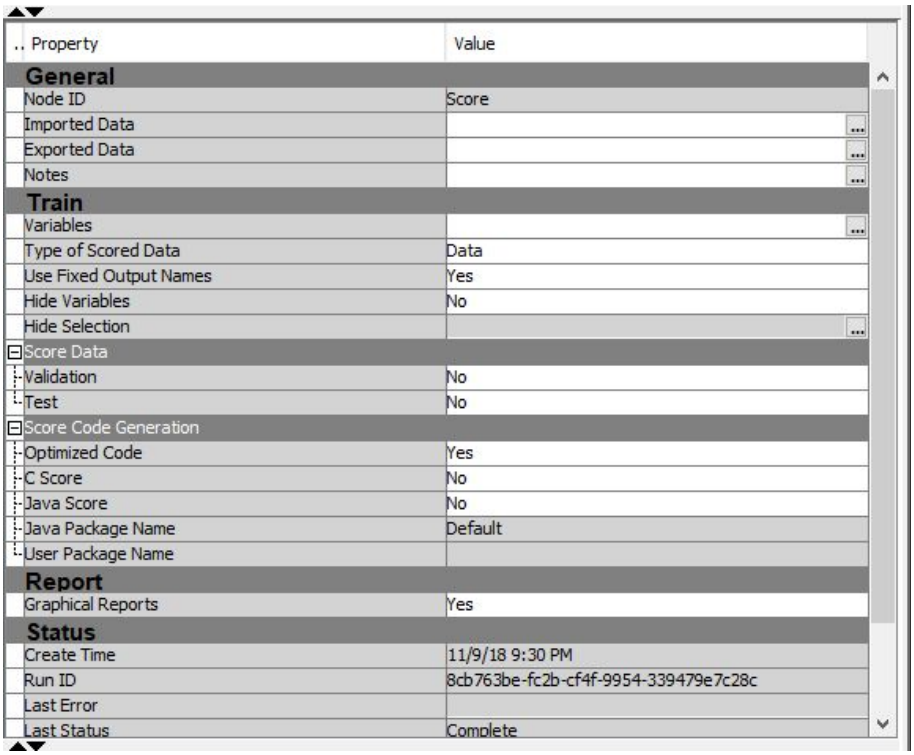


Figure 2: HOSPITAL_READMIT_TO_SCORE_10PCT Metadata



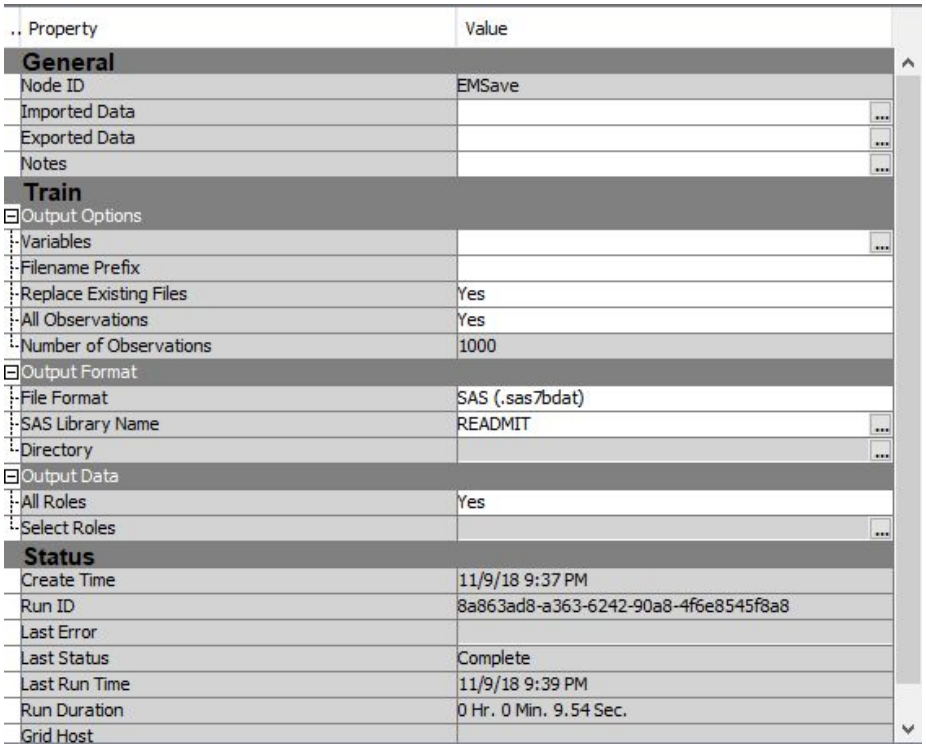
Name	Role	Level
ADMIT_DATE	Time ID	Interval
DISCHARGE_DA	Time ID	Interval
DIAGNOSIS_LO	Text	Nominal
PROCEDURE_LC	Text	Nominal
readmit_number	Rejected	Binary
CITY	Rejected	Nominal
readmit_date	Rejected	Interval
County_name	Rejected	Nominal
DAYS_ICU	Rejected	Unary
ZIP	Rejected	Interval
readmit_discharg	Rejected	Interval
readmit_days	Rejected	Nominal
HOSPITAL	Rejected	Nominal
NUMBER_CHRO	Rejected	Unary
DOCTOR	Input	Interval
LENGTH_OF_ST	Input	Interval
ENCOUNTER_KE	Input	Interval
DISCHARGED_T	Input	Nominal
AllocProportion	Input	Binary
ICU_DAYS	Input	Interval
ActualProportion	Input	Binary
Num_Chronic_C	Input	Nominal
STATECODE	Input	Nominal
PatientAge	Input	Interval
operationcount	Input	Nominal
REGION	Input	Nominal
SampleSize	Input	Binary
Standard_Order	Input	Binary
SamplingWeight	Input	Binary
PATIENT_NUMB	Input	Interval
SelectionProb	Input	Binary
gender	Input	Binary
Y	Input	Interval
Total	Input	Binary
Department	Input	Nominal
Diagnosis_Group	Input	Nominal
op_visits6	Input	Interval
race_cd	Input	Nominal
order_set_used	Input	Binary

Figure 3: Score Node Properties



Property	Value
General	
Node ID	Score
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Type of Scored Data	Data
Use Fixed Output Names	Yes
Hide Variables	No
Hide Selection	
Score Data	
Validation	No
Test	No
Score Code Generation	
Optimized Code	Yes
C Score	No
Java Score	No
Java Package Name	Default
User Package Name	
Report	
Graphical Reports	Yes
Status	
Create Time	11/9/18 9:30 PM
Run ID	8cb763be-fc2b-cf4f-9954-339479e7c28c
Last Error	
Last Status	Complete

Figure 4: Save Data Node Properties



Property	Value
General	
Node ID	EMSave
Imported Data	
Exported Data	
Notes	
Train	
Output Options	
Variables	
Filename Prefix	
Replace Existing Files	Yes
All Observations	Yes
Number of Observations	1000
Output Format	
File Format	SAS (.sas7bdat)
SAS Library Name	READMIT
Directory	
Output Data	
All Roles	Yes
Select Roles	
Status	
Create Time	11/9/18 9:37 PM
Run ID	8a863ad8-a363-6242-90a8-4f6e8545f8a8
Last Error	
Last Status	Complete
Last Run Time	11/9/18 9:39 PM
Run Duration	0 Hr. 0 Min. 9.54 Sec.
Grid Host	

Figure 5: Snapshot of Scored Dataset Probabilities

	d: readmit_number=0	Validated: readmit_number=1	Unnormalized Into: readmit_number	b_readmit_number	Node	Probability of Classification	Probability for level 1 of readmit_number	Prediction for readmit_number
1		1.0	1.0	1.0	31.0	1.0	1.0	1
2		1.0	1.0	1.0	27.0	1.0	1.0	1
3		1.0	1.0	1.0	31.0	1.0	1.0	1
4		1.0	1.0	1.0	31.0	1.0	1.0	1
5		1.0	1.0	1.0	27.0	1.0	1.0	1
6		1.0	1.0	1.0	27.0	1.0	1.0	1
7		1.0	1.0	1.0	31.0	1.0	1.0	1
8	92156862	0.99019607843137	1.0	1.0	61.0	1.0	1.0	1
9		1.0	1.0	1.0	27.0	1.0	1.0	1
10		1.0	1.0	1.0	31.0	1.0	1.0	1
11		1.0	1.0	1.0	31.0	1.0	1.0	1
12		1.0	1.0	1.0	24.0	1.0	1.0	1
13		1.0	1.0	1.0	31.0	1.0	1.0	1
14		1.0	1.0	1.0	31.0	1.0	1.0	1
15		1.0	1.0	1.0	31.0	1.0	1.0	1
16		1.0	1.0	1.0	31.0	1.0	1.0	1
17		1.0	1.0	1.0	31.0	1.0	1.0	1
18		1.0	1.0	1.0	27.0	1.0	1.0	1
19		1.0	1.0	1.0	31.0	1.0	1.0	1
20		1.0	1.0	1.0	27.0	1.0	1.0	1
21		1.0	1.0	1.0	27.0	1.0	1.0	1
22		1.0	1.0	1.0	24.0	1.0	1.0	1
23		1.0	1.0	1.0	27.0	1.0	1.0	1
24		1.0	1.0	1.0	31.0	1.0	1.0	1
25		1.0	1.0	1.0	31.0	1.0	1.0	1
26		1.0	1.0	1.0	31.0	1.0	1.0	1
27		1.0	1.0	1.0	31.0	1.0	1.0	1
28		1.0	1.0	1.0	24.0	1.0	1.0	1
29		1.0	1.0	1.0	31.0	1.0	1.0	1
30		1.0	1.0	1.0	31.0	1.0	1.0	1
31		1.0	1.0	1.0	31.0	1.0	1.0	1
32		1.0	1.0	1.0	31.0	1.0	1.0	1
33		1.0	1.0	1.0	31.0	1.0	1.0	1

Figure 6: Distribution of Readmission Probability