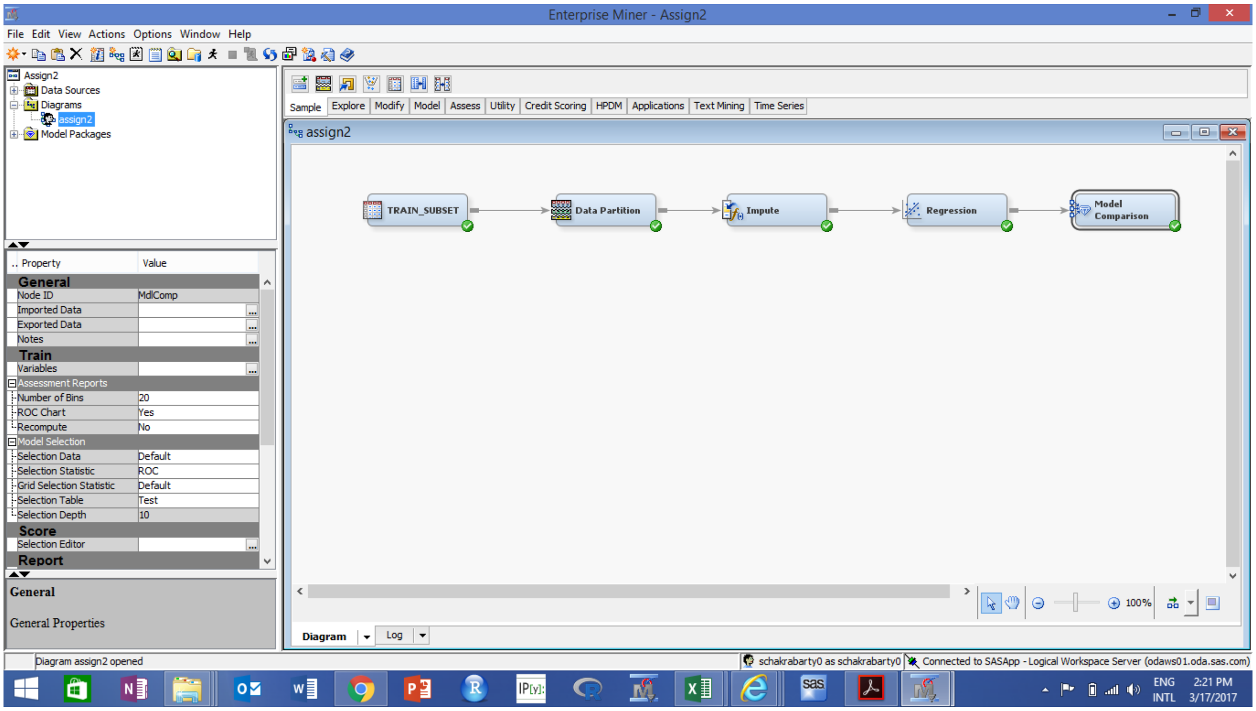
Assignment 2

1. Screenshot of the EM diagram



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	ROC
Grid Selection Statistic	Default
Selection Table	Test
Selection Depth	10
Score	
Selection Editor	
Report	

General

General Properties

Diagram assign2 opened

schakrabarty0 as schakrabarty0

Connected to SASApp - Logical Workspace Server (odaws01.oda.sas.com)

First, we imported the train subset data set on the EM cloud environment. Then we did the following

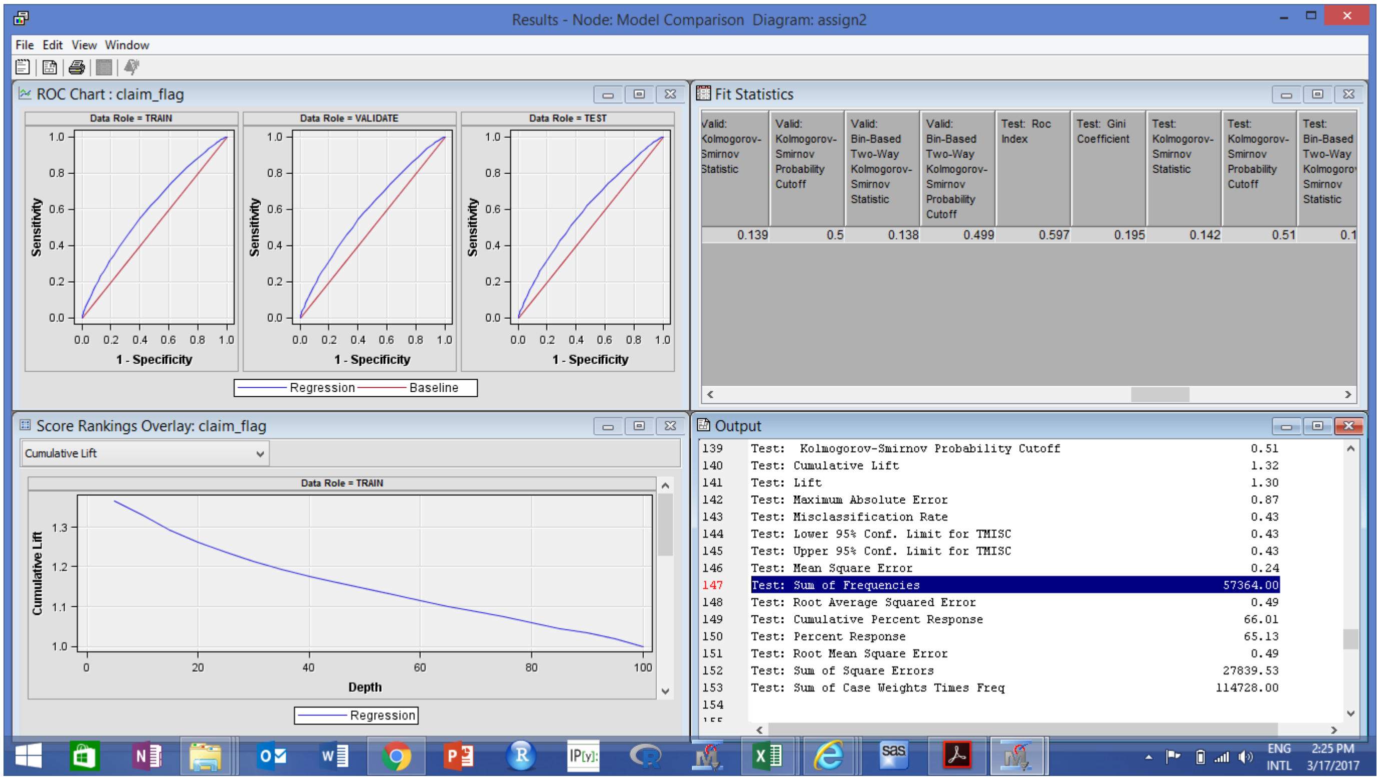
a) To partition the data, we configured the properties of partition component by setting up data set allocations as training = 40; validation = 30; test = 30

b) To impute, we configured the properties of impute component by setting up class variables default input method = count

c) Added regression component and set up the class target regression type as logistic regression; selection model as forward and selection criterion as validation error

d) To get the test AUC value we then added the model comparison component

2. Test AUC = 0.597; No of observations in the test set = 57364



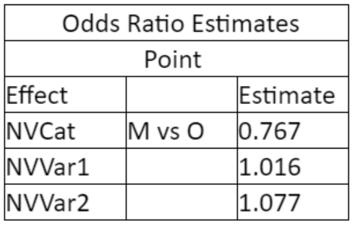
Valid: Kolmogorov-Smirnov Statistic	Valid: Kolmogorov-Smirnov Probability Cutoff	Valid: Bin-Based Two-Way Kolmogorov-Smirnov Statistic	Valid: Bin-Based Two-Way Kolmogorov-Smirnov Probability Cutoff	Test: Roc Index	Test: Gini Coefficient	Test: Kolmogorov-Smirnov Statistic	Test: Kolmogorov-Smirnov Probability Cutoff	Test: Bin-Based Two-Way Kolmogoro Smirnov Statistic
0.139	0.5	0.138	0.499	0.597	0.195	0.142	0.51	0.1

```
Test:  Kolmogorov-Smirnov Probability Cutoff                   0.51
Test: Cumulative Lift                                          1.32
Test: Lift                                                     1.30
Test: Maximum Absolute Error                                   0.87
Test: Misclassification Rate                                   0.43
Test: Lower 95% Conf. Limit for TMISC                          0.43
Test: Upper 95% Conf. Limit for TMISC                          0.43
Test: Mean Square Error                                        0.24
Test: Sum of Frequencies                                   57364.00
Test: Root Average Squared Error                               0.49
Test: Cumulative Percent Response                             66.01
Test: Percent Response                                        65.13
Test: Root Mean Square Error                                   0.49
Test: Sum of Square Errors                                 27839.53
Test: Sum of Case Weights Times Freq                      114728.00
```

3. Most important variable in the model in terms of the odds that a policy will have a claim associated with it

Based on the Chi square value (>100), we have selected three variables as attached in the following:

Analysis of Maximum Likelihood Estimates								
Standard Wald Standardized								
Parameter		DF	Estimate	Error	Chi-Square	Pr > ChiSq	Estimate	Exp(Est)
NVCat	M	1	-0.2653	0.0196	182.67	<.0001		0.767
NVVar2		1	0.0738	0.00728	102.76	<.0001	0.0476	1.077
NVVar3		1	0.0908	0.00740	150.60	<.0001	0.0591	1.095



Odds Ratio Estimates

		Point
Effect		Estimate
NVCat	M vs O	0.767
NVVar1		1.016
NVVar2		1.077

Explanation:
NVCat M: Holding all other variables constant, NVCat being M changes the odds of the claim event occurring by a factor of 0.767 over the reference level (0) on average

NVVar2: Holding all other variables constant, for one unit increase in NVVar2, the odds of claim event occurring changes by a factor of 1.077 on average

NVVar3: Holding all other variables constant, for one unit increase in NVVar3, the odds of claim event occurring changes by a factor of 1.095 on average