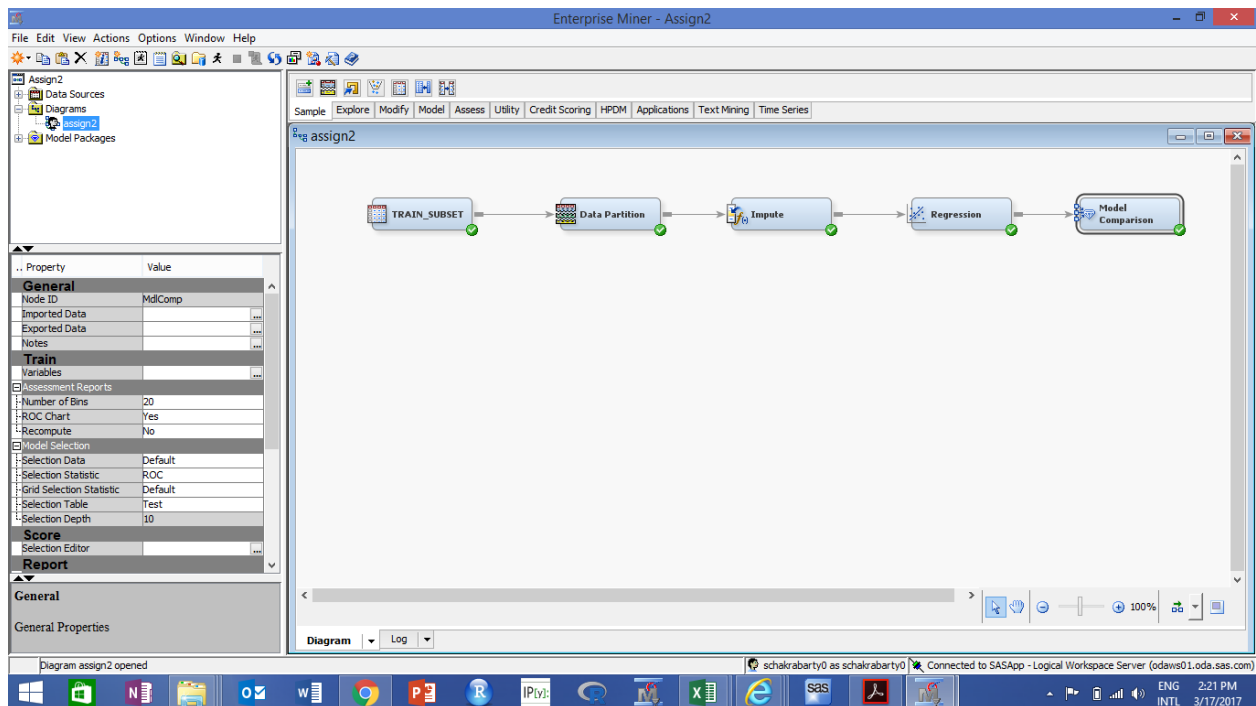


Assignment 2

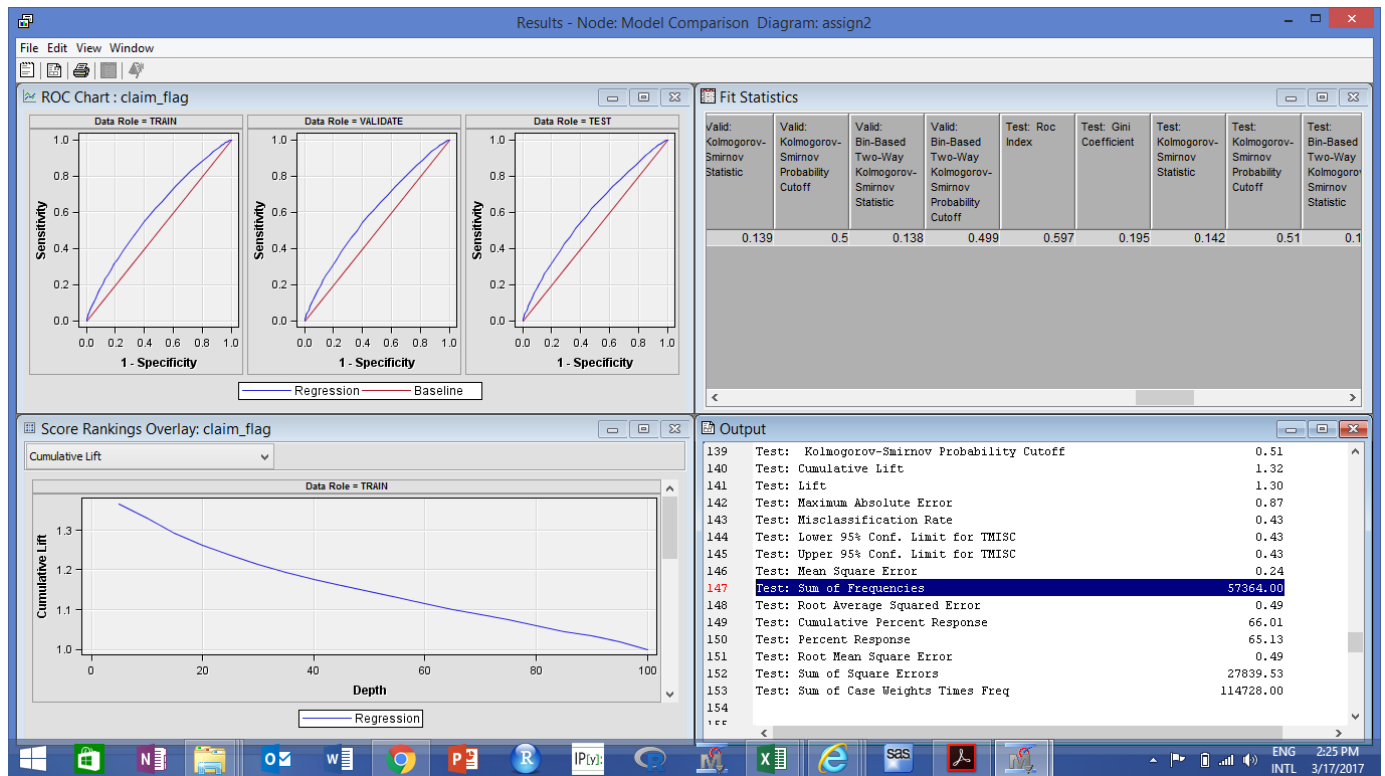
1. Screenshot of the EM diagram



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

- To partition the data, we configured the properties of partition component by setting up data set allocations as training = 40; validation = 30; test = 30
- To impute, we configured the properties of impute component by setting up class variables default input method = count
- Added regression component and set up the class target regression type as logistic regression; selection model as forward and selection criterion as validation error
- 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



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